Universidad Carlos III de Madrid

Escuela Politécnica Superior

Grado en Ingeniería de Sistemas Audiovisuales

TRABAJO FIN DE GRADO

TOWARDS MULTI-MODAL AFFECTIVE SCENE RECOGNITION IN VIDEO-CLIPS

Autor: Gonzalo Solana Pascual

Tutor: Fernando Fernández Martínez

Madrid, 27 de Septiembre de 2015

RESUMEN

Internet, utilizado como una plataforma digital de información ha sido el promotor de una revolución tecnológica histórica, además de contener un gran conjunto diverso de información multimedia. El estudio objetivo de cómo afecta este contenido multimedia a los usuarios está creciendo de manera exponencial, ya que el potencial que tienen las aplicaciones desarrolladas en este campo es muy alto. Esto hace que la experiencia del usuario al ver un vídeo sea cada vez más personalizada con respecto al contenido que visualiza, o que sistemas de seguridad basados en eventos multimedia sean más robustos.

En este Trabajo Fin de Grado se estudia la influencia del audio en la respuesta afectiva del espectador al visualizar un vídeo. Para ello se emplea la base de datos LIRIS-ACCEDE, la cual contiene 9.800 extractos de vídeos y es de reciente lanzamiento a la comunidad científica. Una vez analizada la base de datos se extrae el archivo de audio perteneciente a cada uno de los vídeos, de esta forma es posible trabajar sólo con el audio. Con esto, se definen los objetivos del proyecto: (i) extracción de características acústicas de bajo nivel, las cuales derivan del resultado de un procesado básico de la señal de audio, y (ii) proposición y extracción de características acústicas de alto nivel, las cuales vienen precedidas de un proceso de segmentación del fichero de audio en varios eventos acústicos tales como “voz”, “música”, “voz sobre música”, etc. A esta investigación le sigue un proceso de experimentación, en el que se evaluará la calidad de las características acústicas (de alto y bajo nivel) extraídas y propuestas. Tras este último paso podremos concluir si dichas características cumplen con el objetivo inicial de modelar la respuesta afectiva de un espectador al visualizar un vídeo, o por el contrario no influyen en la respuesta afectiva del usuario.

Palabras clave: audio information retrieval, affective response, emotion recognition, audio feature extraction.

CONTENIDO

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE TABLAS

CAPÍTULO 1. INTRODUCTION

1.1 CONTEXT AND OBJECTIVES

The recent technological progress that has become evident during the last decade has provided a digital platform that has allowed users and professionals to access information and manage data.

The internet, used as a digital platform for information, has been the promoter of a historical revolution of technology. It has become a vessel for a vast and diverse ensemble of multimedia content. However, the Internet's expansion and exponential growth aggravates the problem of the enormous amount of information that cannot be easily accessed, classified and indexed.

In the case of audiovisual information, Internet permits the access to large databases that contain a huge volume of information, including images, videos and audio files.

Every day, millions of users are able to watch, upload or download videos using different platforms. One example could be YouTube, which has more than a thousand million users and where hundreds of millions of hours of video are watched daily. It generates thousands of millions of video views and each minute 300 hours of video are being uploaded [37].

These videos are composed by the audio and a set of images that follow each other every few milliseconds. Only with the study of the characteristics of the audio, it is possible to obtain useful and varied information. Henceforth, the research focused in this field of study has proved its many applications. Returning to the previous example about YouTube, the results of the research can facilitate the access to information or reduce the time invested in searching for a video.

According to this, it is very interesting a recent study about how the characteristics of audio can affect the perceptual response of the spectator while watching a video.

The main problem that arises in the approach is that the affective perception of human beings is mostly subjective. Thus, a great amount of reliable, well-structured and labeled data is needed in order to draw objective results and conclusions capable of modeling people's emotions.

The starting point of this study is the launch of the LIRIS-ACCEDE video database [1]. This database is addressed to the investigations related to the affective response of the spectators. It is composed by 9,800 fragments whose duration goes from 8 to 12 seconds each, and they belong to 160 movies of diverse origin and theme. In addition, the videos are labeled by means of crowdsourcing in a two-dimension space: *valence* and *arousal*. These labels are related to the emotion that a multimedia event evokes in a person: *valence* is the type of emotion (sadness, happiness) and *arousal* is the intensity with which that emotion is experienced (inactive, excited). It is because of these labels that the feelings that are induced in a user while watching the videos can be accurately measured.

With all the information provided by the LIRIS-ACCEDE database, a large set of high quality labeled data is available to continue with the study.

Once obtained the data set that is needed to work and to compare the experiments' results, it is necessary to extract the audios of every video.

Each audio will be analyzed and treated in two different ways to obtain two combinations of acoustic characteristics: low level and high level characteristics.

Low level characteristics are those that correspond to the result of a basic processing of the audio signal through MIRtoolbox [2] (spectrum, autocorrelation, standard deviation, energy of the signal, zero crossings, etc.).

High level acoustic characteristics are possible to obtain from an audio file once this file has undergone a process that segments it into acoustic events. Some acoustic events would be: "music", "voice", "voice over music", "others" etc. An example of these high level characteristics could be checking if an audio fragment contains music or if it is a person talking.

Bearing this in mind, a project has been developed with the following main objectives:

- To obtain high and low level acoustic characteristics. For this purpose, we will use tools to process audio such as MIRtoolbox [2] as well as tools capable of segmenting audio fragments in various acoustic events.

- To propose new acoustic characteristics of high level based on the results obtained through segmenting tools.

- To evaluate high and low level acoustic characteristics, both in an independent and combined way, and to compare the results with the labeled videos in the LIRIS-ACCEDE database [1].

1.2 DOCUMENT STRUCTURE

In this section, the arrangement of the document is introduced. It is divided in different chapters, with the aim of providing a clear explanation of the investigation and the development of the project.

- **Chapter 1**: Introduction to the context and motivations of the investigation, together with a brief description of the main objectives.

- **Chapter 2**: Analysis of the state of the art and applications of this technology. Requisites, restrictions and regulatory framework of the tools and equipment used during the project.

- **Chapter 3**: Implementation of the technical solution. Presentation of general concepts and description of the database. Obtaining of low level characteristics and proposal of high level characteristics.

- **Chapter 4**: Development of the experimentation performed in the project. Machine Learning concepts and software employed: WEKA.

- **Chapter 5**: Presentation and analysis of the results obtained during the experimentation phase.

- **Chapter 6**: Diagram, organization and budget of the project

- **Chapter 7**: Main conclusions and future lines of work.

CAPÍTULO 2. PLANTEAMIENTO DEL PROBLEMA: ESTADO DEL ARTE

2.1 RESPUESTA AFECTIVA DE UN ESPECTADOR AL VISUALIZAR UN VÍDEO

El contenido afectivo de un vídeo puede ser definido como la intensidad y el tipo de sentimiento o emoción (ambas son denominadas como *afecto*) que son esperadas como reacción del usuario al visualizar el vídeo.

Esta "esperada" emoción puede ser vista como la que el director (o creador) del vídeo ha querido mostrar a la audiencia que visualiza el vídeo o como la emoción que se ha suscitado a partir de la mayoría de la audiencia que ha visto el vídeo.

La capacidad de extraer automáticamente el contenido de vídeo de esta naturaleza conduce a un alto nivel de personalización a la hora de transmitir vídeo a usuarios privados, así como a ampliar considerablemente las posibilidades de manipulación y presentación eficaz de grandes cantidades de material audiovisual almacenado en bases de datos de vídeos.

La importancia de la ampliación de la investigación en el campo del análisis del contenido afectivo de un vídeo, permite un número de nuevas o mejoradas aplicaciones, tales como la indexación y recuperación de vídeo o la entrega de vídeo personalizado. Por ejemplo: obtener todos los vídeos de una base de datos que contengan sólo música, u obtener un ranking de los 10 vídeos que más miedo causen.

2.1.1 MODELOS COMPUTACIONALES DE LA EMOCIÓN

El trabajo sobre el análisis del vídeo afectivo puede ser categorizado en dos subgrupos: análisis continuo del contenido afectivo del vídeo, el cual estima una puntuación afectiva para cada frame del vídeo, y análisis discreto del contenido afectivo del vídeo, el cual asigna una puntuación afectiva a un segmento del vídeo o a la totalidad del mismo.

Los pioneros en el análisis del contenido afectivo en un vídeo fueron Hanjalic y Xu, en su estudio [3] mapearon directamente las características del vídeo en un espacio de dos dimensiones *valencia-arousal* para crear continuas representaciones. Finalmente ofrecieron una evaluación cualitativa de su modelo. Malandrakis *et al.* también propusieron en [4] un análisis continuo del contenido afectivo de un vídeo a partir de características extraídas de

cada frame perteneciente al vídeo, y presentado a dos Modelos Ocultos de Markov (HMMs). Cada clasificador se encargaba de modelar una variable diferente (*valencia* o *arousal*) de forma independiente a la otra.

El análisis discreto del contenido afectivo en un vídeo ha sido más frecuentemente investigado durante la última década. El primer autor en proponer un modelo donde los clasificadores son adoptados para análisis afectivos fue Kang [5]. En este estudio sugirió la detección de estados afectivos en películas incluyendo "tristeza", "felicidad" y "miedo" a través de características de bajo nivel usando HMMs. Siguiendo el mismo camino, Wang y Cheong [6] introdujeron características inspiradas en psicología y en normas de rodaje de películas. Sun and Yu [7] también desarrollaron un estudio en este campo, en el que las unidades de vídeo eran primero representadas en diferentes granularidades usando una curva de excitación basada en la curva de *arousal* presentada por Hanjalic y Xu en [3]. Después, cuatro HMMs eran entrenados independientemente usando características extraídas de las granularidades mencionadas anteriormente para reconocer uno de entre cuatro estados emocionales: "alegría", "enfado", "tristeza" y "miedo".

En esta investigación vamos a tratar con el análisis discreto del contenido afectivo en un vídeo. El principal motivo viene dado porque el proyecto abordará únicamente el estudio de la influencia del audio en la respuesta afectiva del espectador. La alternativa basada en un análisis continuo queda propuesta como una posible línea futura para el presente proyecto de investigación.

2.2 ESTADO ACTUAL DE LA TECNOLOGÍA DE SEGMENTACIÓN DE ARCHIVOS DE AUDIO

El problema de la distinción de señales de voz de otras señales de audio (p.ej., música) se ha hecho cada vez más importante como los sistemas de reconocimiento automático de voz, los cuales son aplicados a dominios multimedia del mundo real, como la transcripción automática de noticias, en las cuales la voz típicamente está intercalada con segmentos de música y otros ruidos de fondo. Una etapa de pre-procesado que segmente la señal en periodos de voz y de no-voz es muy importante para mejorar la precisión del reconocimiento.

Por otro lado, detectar automáticamente partes de música en señales de audio procedentes de televisión o de radio se está convirtiendo en una importante tarea para poder resolver el actual incremento de la demanda de sistemas de indexado multimedia y sistemas de control de copyright. En tales señales de audio, la música a menudo es superpuesta por la narración, conversación, u otros sonidos ambientales.

Trabajos previos en la segmentación de voz y música se han centrado en el análisis de características o en la arquitectura del sistema.

Sobre el análisis de las características, podemos destacar [11] donde los autores combinan *Mel Frequency Cepstral Coefficients* (MFCCs) con otras características como modulación de la energía a 4Hz, frames con bajo porcentaje de la energía, centroide espectral, punto de roll-off espectral, flujo espectral, ratio de cruces por cero y borde espectral. En [12] las características basadas en la ecualización del histograma son propuestas para discriminar voz y música. En [13] una red neuronal artificial (ANN) entrenada únicamente con voz limpia es usada como modelo canal en la salida, cuya entropía y "dinamismo" son medidos cada 10 ms. Estas características son entonces integradas en tiempo a través de un Modelo oculto de Markov (HMM) ergódico de dos estados con una duración límite mínima en cada estado. Finalmente, en [14], para la discriminación entre voz y música se emplearon características como el error cuadrático medio y los cruces por cero de la señal de audio.

Con respecto a la arquitectura del sistema, en [15] se propone un algoritmo de decisión basado en una estructura de tipo árbol para la segmentación de voz y música. [16] presenta una comparación entre dos técnicas diferentes para la discriminación entre voz y música. El primer método está basado en el ratio de cruces por cero de la señal y en una clasificación bayesiana. El segundo método usa más características y está basado en redes neuronales (específicamente en un perceptrón multicapa).

En el trabajo [17], los autores proponen un sistema jerárquico de segmentación de archivos de audio pertenecientes a noticias. Finalmente, en [18] es propuesto e implementado un sistema capaz de segmentar la señal de audio diferenciando entre voz y música usando una probabilidad posterior basada en características. Este sistema usa Modelos ocultos de Markov basados en modelos acústicos que son entrenados para posteriores cálculos probabilísticos.

En este proyecto se hará uso de la herramienta de segmentación proporcionada por A. Gallardo y R. San Segundo en [8], la cual fue creada como propuesta al concurso de segmentación de audio Albayzin 2010 [9] donde obtuvo el primer puesto de la evaluación. El sistema empleado está basado en Modelos Ocultos de Markov (HMMs), incluyendo un HMM de 3 estados por cada clase acústica ("voz", "música", "voz con ruido de fondo" y "voz con música de fondo"). Se implementaron dos arquitecturas: la primera corresponde a un sistema *one-step,* la cual obtuvo un 25.3% de error medio, y la segunda corresponde a un sistema jerárquico que obtuvo un error medio de 23.9%. Para el desarrollo de esta investigación se hará uso del sistema *one-step*.

2.2.1 APLICACIONES

Las aplicaciones que tienen la segmentación en eventos acústicos y el estudio de las señales de audio abarcan multitud de campos, desde sistemas de seguridad basados en reconocimiento de voz hasta la entrega de contenidos personalizados a los usuarios.

Una aplicación que está en continuo desarrollo y mejora son los sistemas de control de copyright. En ellos es muy importante la semejanza entre el contenido multimedia que aparece en el nuevo material y el contenido multimedia ya almacenado en una base de datos. Ambos contenidos son comparados para determinar si se están infringiendo las normativas de propiedad de los usuarios. Esta herramienta de comparación es muy útil si funciona bien, ya que automatiza un proceso muy costoso que una persona tardaría mucho más tiempo en desarrollar. Por ejemplo, en YouTube, cada vez que un usuario sube un vídeo a la plataforma es comparado con todas la referencias a vídeos que hay en la base de datos. Las referencias son de varios tipos: audio, vídeo, personas que aparecen en dicho vídeo, etc. Una vez comparado el vídeo subido con todas las referencias almacenadas, el vídeo puede ser bloqueado o liberado. Con la cantidad de contenido multimedia que es subido cada hora a YouTube, serían necesarias 36.000 personas trabajando 24 horas todos los días para comprobar si se abusa de copia en el contenido de cada vídeo [19].

Otras dos aplicaciones que van prácticamente unidas de la mano son el reconocimiento de voz y la transcripción de audio. Actualmente todos los dispositivos móviles disponen de una herramienta con la que damos órdenes al teléfono móvil y ésta es capaz de ejecutar dicha operación sobre el dispositivo. Los pasos que usa esta herramienta son: (i) Transcripción de la voz humana usando reconocimiento automático de voz, (ii) Uso del procesado natural del lenguaje para traducir el texto *reconocido* por el dispositivo a texto que el teléfono es capaz de comprender, y (iii) Análisis del texto parseado relacionándolo con las herramientas que dispone el dispositivo [20].

Por último, una aplicación que conlleva un negocio de trasfondo, es la entrega de vídeo personalizado. Una vez analizado el contenido afectivo de un vídeo, y aplicado a cualquier evento multimedia que un espectador visualiza, seremos capaces de saber qué vídeos son del agrado o desagrado de una persona. Gracias a esto, plataformas como YouTube son capaces de ofrecer nuevos modelos alternativos a los esquemas clásicos basados en texto y/o metadatos; estos nuevos modelos analizan al completo el contenido multimedia del vídeo, audio e imágenes, extrayendo información relevante de los mismos. Esto es muy útil para los sistemas de recomendación e indexado, que gracias a dicha información ampliada son capaces de ofrecer mejores prestaciones tanto de recomendación como de búsqueda.

Como se aprecia, es importante el estudio de la respuesta afectiva de un espectador al visualizar un vídeo ya que tiene variedad de aplicaciones y utilidades que pueden facilitarnos el uso de algunas herramientas y mejorar la vida cotidiana de las personas.

2.3 MOTIVACIONES Y PROPÓSITOS

Se ha analizado la trascendencia que el análisis de la respuesta afectiva de un espectador al visualizar un vídeo puede tener en la vida cotidiana, así como los fines económicos que este campo puede abarcar.

Es por eso, que una vez dispuestos todos los recursos, resulta de gran interés el estudio que se va a llevar a cabo en este proyecto.

En primer lugar, el análisis y el uso de la base de datos LIRIS-ACCEDE [1] es de especial interés para la comunidad científica al ser un proyecto lanzado recientemente, el cual dispone de una gran cantidad de vídeos etiquetados fielmente mediante crowdsourcing sobre los que se pueden realizar profundas investigaciones relacionadas con el ámbito de la respuesta afectiva.

En segundo lugar, la disposición de la herramienta MIRtoolbox [2] hace que sea posible la extracción de aproximadamente 400 características de bajo nivel de la señal de audio y por tanto la hace muy útil para ver cómo estas características a nivel de procesado de señal afectan a la respuesta afectiva de un espectador.

Por último, la disposición a nuestro alcance de la herramienta de segmentación empleada por A. Gallardo y R. San Segundo en [8] hace posible segmentar los archivos de audio en diferentes eventos acústicos y así proponer nuevas características acústicas de alto nivel que nos ayuden a entender parcialmente cómo las personas reaccionamos al visualizar un vídeo. Decimos parcialmente ya que sólo empleamos el audio del vídeo en esta investigación; lo cual no quiere decir que las imágenes que componen el vídeo no sean importantes o incluso fundamentales para resolver el problema de forma completa. Es por ello, que el estudio de las imágenes del vídeo se presenta como como una línea futura, así como la investigación de la conveniente combinación de ambos tipos de características: visuales y acústicas.

2.4 REQUISITOS

A continuación se muestran los requisitos mínimos necesarios para llevar a cabo el siguiente proyecto:

- Equipo informático que cumpla con los requisitos mínimos exigidos por la versión de Matlab R2015b [21] (ya que es el software de entre todos los empleados en el proyecto con requisitos mínimos más altos). Estos requisitos mínimos son: Sistema operativo Windows 7 en adelante, Mac OS X 10.9.5 en adelante, o distribuciones cualificadas de Linux especificadas en la página web de Mathworks. Cualquier procesador Intel o AMD x86 capaz de soportar el set de instrucciones SSE2. 2 GB de memoria RAM, siendo recomendable disponer de 4 GB. Espacio en el disco duro suficiente para contener Matlab con las herramientas adecuadas, más la colección de videos sobre la que trabajar (aproximadamente 15GB).
- Base de datos LIRIS-ACCEDE [1], la cual contiene 9.800 vídeos sobre los que trabajaremos y contrastaremos resultados.
- El software necesario será: MATLAB, MIRToolbox [2], OpenSmile [22], SoX [26], HTK Speech Recognition Toolkit [23], Python(x,y) [24] y WEKA [25]. Del software enumerado anteriormente solo se necesitará licencia para MATLAB y HTK, siendo gratuita para HTK. El resto de programas son de libre distribución.

2.5 RESTRICCIONES Y MARCO REGULADOR

Como hemos comentado anteriormente, el software MIRtoolbox, OpenSmile , Python(x,y), SoX y WEKA son de distribución gratuita y por tanto se puede utilizar sin restricciones.

En el caso de MATLAB, se requiere una licencia original para la revisión R2015b en adelante.

Con respecto a la herramienta HTK Speech Recognition Toolkit, es necesario aceptar una licencia disponible en su página web [23] previamente a la descarga del software. Una vez que aceptemos esta licencia, HTK estará disponible para nuestra investigación de forma gratuita.

También se ha dispuesto de una herramienta de segmentación [8], la cual fue propuesta para un concurso/evaluación de segmentación de audios denominado Albayzin-2010 [9]. Los modelos empleados por esta herramienta han sido entrenados con una base de datos de audios pertenecientes a un noticiero de un canal de televisión catalana, los cuales fueron grabados por el Centro de Investigación TALP perteneciente a la UPC [10].

CAPÍTULO 3. DISEÑO E IMPLEMENTACIÓN DE LA SOLUCIÓN TÉCNICA

El objetivo de esta sección es mostrar el desarrollo tanto técnico como teórico de la implementación llevada a cabo para desarrollar el proyecto.

La investigación que se lleva a cabo tiene como objetivo el estudio de la influencia del audio en la respuesta afectiva de un espectador al visualizar un vídeo. Para ello se estudiarán las características acústicas de bajo nivel proporcionadas por MIRtoolbox [2], así como las que más peso tienen a la hora de modelar la respuesta afectiva del espectador. Por otro lado se propondrán características acústicas de alto nivel derivadas de la segmentación de ficheros de audio, y se estudiará cómo éstas influyen en la persona que visualiza el vídeo.

Con esto, sabremos si las características acústicas de bajo nivel existentes y las nuevas de alto nivel propuestas modelan significativamente la respuesta afectiva del espectador o por el contrario podríamos descartar este tipo de características al no ser relevantes.

La sección comienza con una descripción de conceptos generales que son usados frecuentemente a lo largo del proyecto y son considerados de importancia para este estudio, seguido de un esquema general del desarrollo técnico. Más adelante se profundiza en un apartado sobre la base de datos empleada, donde se explicarán los fundamentos seguidos para su elección. Y por último habrá dos apartados, donde cada uno se centrará en los dos tipos de características acústicas que se estudian: características acústicas de bajo nivel y características acústicas de alto nivel.

3.1 CONCEPTOS GENERALES

3.1.1 AROUSAL

El concepto de *arousal* ha sido definido de diferentes formas: "estado fisiológico de vigilancia y anticipación que prepara el cuerpo para la acción", "estado de preparación para la realización de algo que ayuda a motivar a los realizadores" o "el estado enérgico, o la preparación para la acción que motiva a un realizador a comportarse de una determinada manera".

Para la representación del contenido afectivo, el *arousal* está relacionado con la energía o intensidad de la emoción suscitada en un espectador cuando éste es expuesto a un estímulo, genéricamente hablando (un contenido multimedia en nuestro caso en particular). Su rango es muy subjetivo, pero varía entre excitado o enérgico y tranquilo.

3.1.2 VALENCIA

El concepto de *valencia* también ha sido definido de diferentes formas, pero N. H. Frijda en [32] cerró su definición, reduciéndolo a un término utilizado en psicología, especialmente en el ámbito de las emociones, lo que demuestra el atractivo intrínseco (*valencia* positiva) o la aversión (*valencia* negativa) hacia un evento, objeto o situación.

Para la representación de contenido afectivo, la *valencia* está relacionada con el "tipo" de emoción evocada en el usuario cuando éste es expuesto a un estímulo, genéricamente hablando (un contenido multimedia en nuestro caso en particular). Su rango es bastante subjetivo, pero varía entre algo agradable/positivo y algo desagradable/negativo.

3.1.3 ESCALA VALENCIA-AROUSAL

La escala *valencia-arousal* es un modelo continuo de emoción usada asiduamente en investigaciones relacionadas con el "afecto" o "respuesta afectiva". Fue propuesta por primera vez por Rusell en [36]. El concepto es que cada estado emocional puede ser ubicado en un plano de dos dimensiones con la *valencia* y el *arousal* como ejes. El *arousal* puede variar aquí desde inactivo (p. ej. desinteresado, aburrido) hasta activo (p. ej. excitado), mientras que el margen de la *valencia* va desde lo desagradable (p. ej. triste, estresado) hasta lo agradable (p. ej. feliz, eufórico).

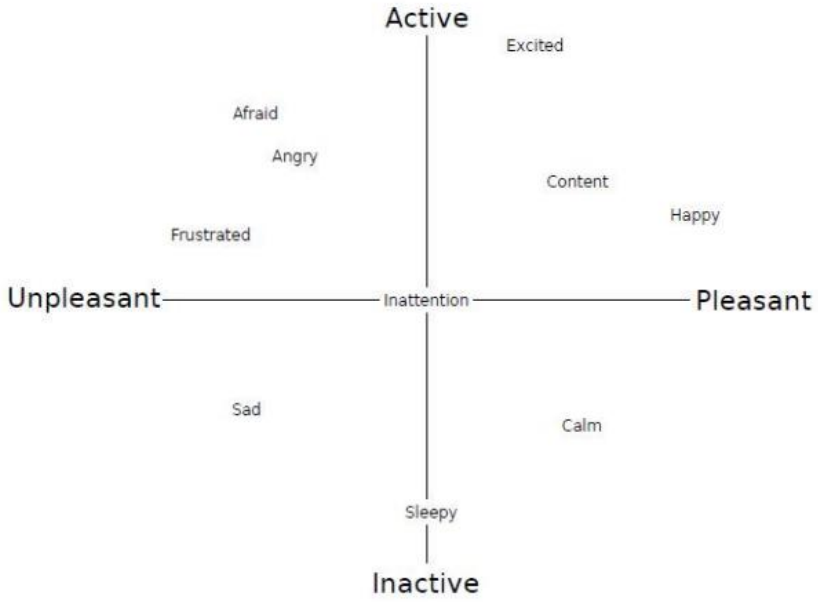


Ilustración 1. Escala de *valencia-arousal*

Mientras que el *arousal* y la *valencia* explican la mayoría de las variaciones en los estados emocionales, es discutible introducir al modelo una tercera dimensión de control o dominancia e incluso una cuarta dimensión que recoja la imprevisibilidad. El rango de la dominancia va desde un sentimiento de debilidad o impotencia (sin control) hasta un sentimiento "todopoderoso" (con todo bajo control). La imprevisibilidad está relacionada con la novedad o espontaneidad de una emoción.

Los modelos continuos de las emociones presentan dificultades para los participantes del experimento a la hora de anotar la emoción sufrida al visualizar un vídeo. Considerando que los modelos continuos propuestos tienen valores comprendidos entre 1 y 100 en las escalas de *valencia* y *arousal*, se precisan herramientas de calificación para facilitar la autoevaluación en las escalas continuas. Tres de estas herramientas son las siguientes:

- Autoevaluación del maniquí [27]. Para cada dimensión de *valencia*, *arousal* y *dominancia* hay una serie de maniquíes que representan los diferentes valores a lo largo de los ejes. Los participantes del experimento pueden seleccionar para cada dimensión el maniquí que mejor se ajusta a la emoción que han sentido. (Ver *Ilustración 2*).

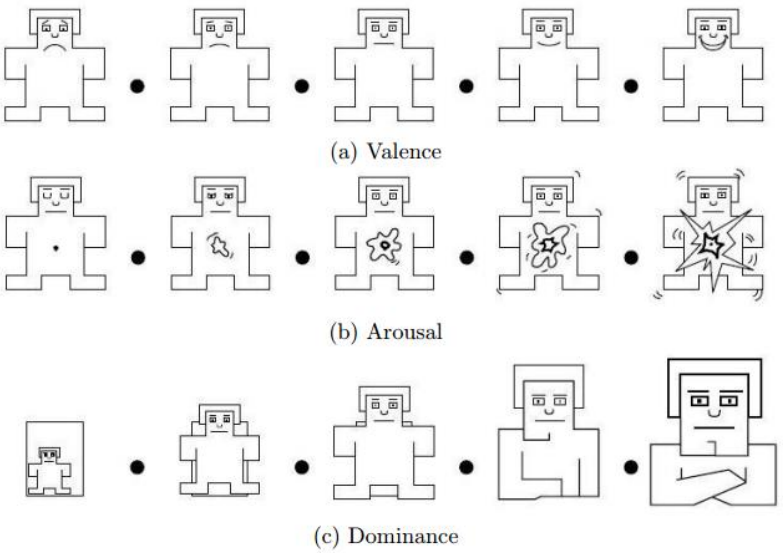


Ilustración 2. Autoevaluación del maniquí

- Rueda de emociones Ginebra [28]. La herramienta permite visualizar el espacio *valencia-arousal* relacionándolo con una serie de palabras emocionales clave que representan los extremos en el espacio. Cada palabra clave es a su vez representada por un conjunto de círculos de diferentes tamaños, donde el tamaño del círculo representa la fuerza de la emoción sentida. Los participantes pueden seleccionar una palabra clave y su correspondiente fuerza eligiendo uno de los círculos. Ha sido demostrado que la elección de la palabra clave/fuerza se correlaciona con la ubicación en el espacio *valencia-arousal* [29]. (Ver *Ilustración 3*).

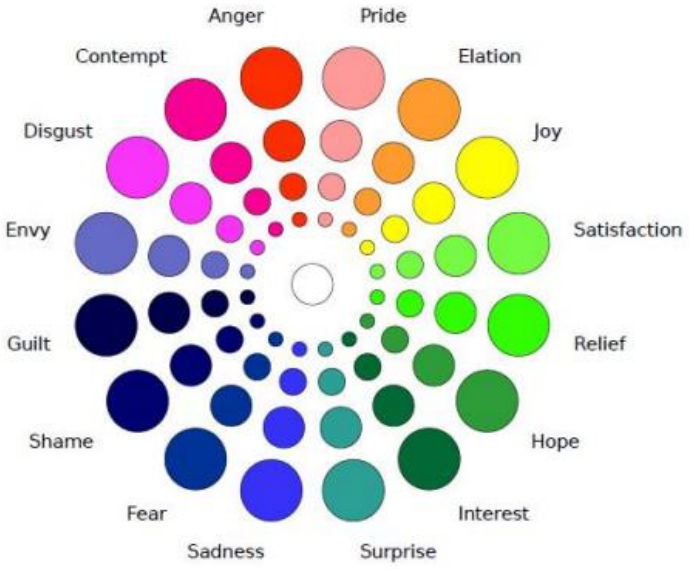


Ilustración 3. Rueda de emociones Ginebra

- Gtrace [30], sucesor de FEELtrace [31]. Esta herramienta permite al usuario grabar en tiempo real la "traza" que especifica la evolución temporal de la emoción sentida a lo largo del tiempo. La herramienta visualiza el vídeo que va a ser anotado y un cursor que es posible mover hacia adelante y hacia atrás a través de la escala diseñada propiamente para el vídeo. (Ver *Ilustración 4*).

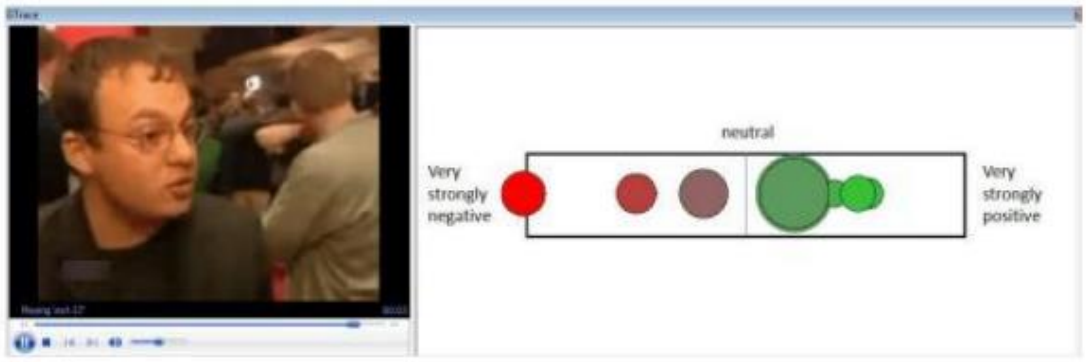


Ilustración 4. Framework de anotación de emociones Gtrace

3.2 ESQUEMA DEL PROYECTO

A continuación se muestra el esquema que se seguirá para desarrollar el proyecto.

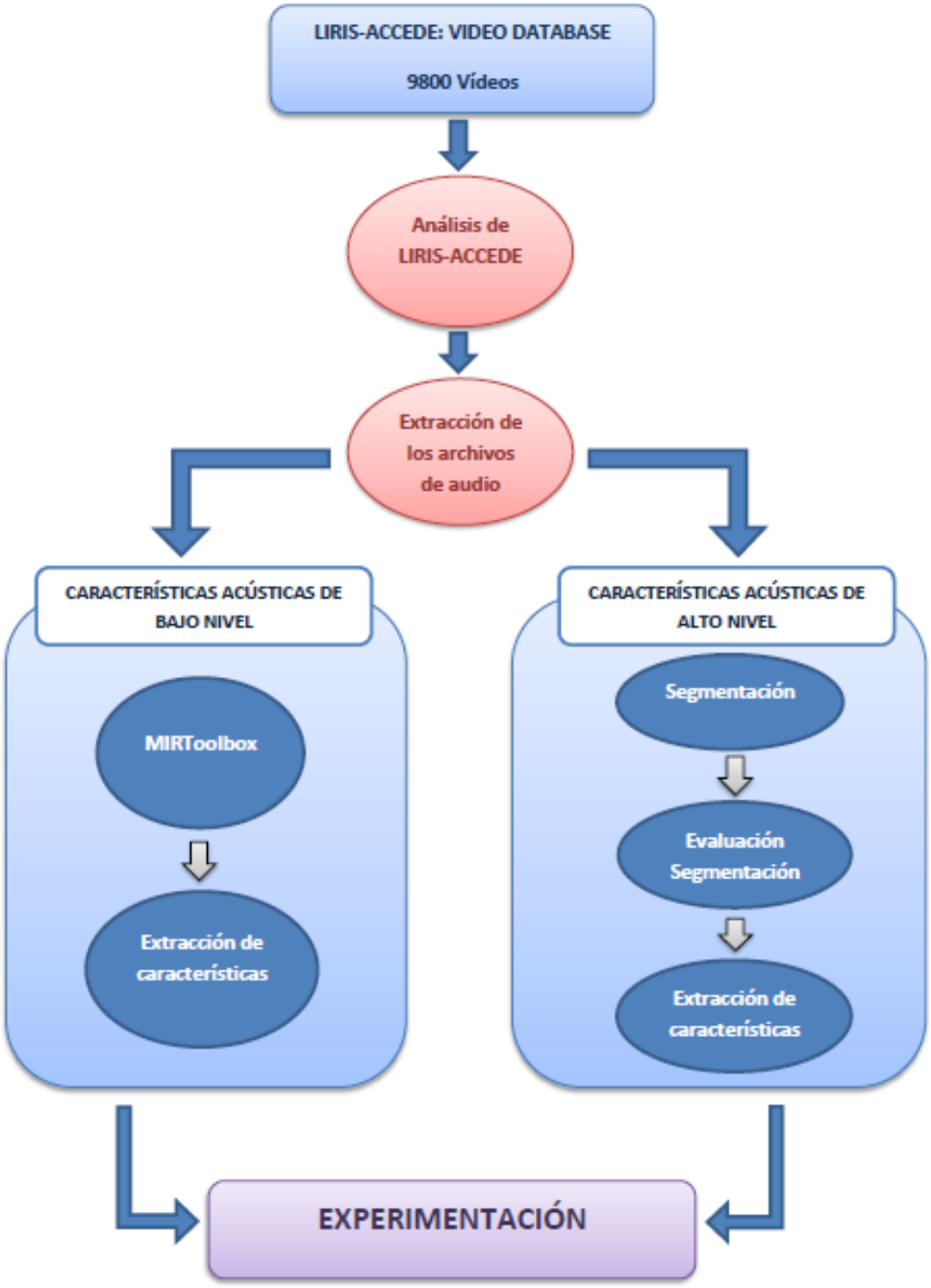


Ilustración 5. Esquema del desarrollo del proyecto

3.3 LIRIS-ACCEDE: VIDEO DATABASE

En este apartado se va a introducir la base de datos de vídeos usada durante el desarrollo del proyecto, así como la composición de dichos vídeos y la justificación de la elección de esta base de datos.

3.3.1 INTRODUCCIÓN

LIRIS-ACCEDE [1], es una base de datos de vídeos realizada con el objetivo de ser empleada en estudios de respuesta afectiva de las personas ante la visualización de un vídeo.

Está compuesta por 9.800 extractos pertenecientes a 160 películas y cortos. Es la mayor base de datos de vídeos etiquetada por una amplia y representativa parte de la población que existe actualmente en la que se usan etiquetas emocionales. Todo estos extractos están categorizados mediante crowdsourcing (colaboración abierta distribuida) en el espacio de dos dimensiones *valencia-arousal*.

LIRIS-ACCEDE está disponible de forma libre para la comunidad investigadora. Por eso, las 160 películas usadas para crear la base de datos son compartidas bajo licencias Creative Commons [33]. Creative Commons es una corporación sin ánimo de lucro que permite usar y compartir tanto la creatividad como el conocimiento a través de una serie de instrumentos jurídicos de carácter gratuito.

Estas películas están clasificadas según 9 géneros representativos que aseguran diversidad en su contenido: Comedia, Animación, Acción, Aventura, Suspense, Documental, Romance, Drama y Miedo.

El idioma principal de los vídeos es el inglés, aunque también se emplea un pequeño conjunto de varios idiomas: francés, alemán, hindú, español, italiano, etc.

3.3.2 COMPOSICIÓN

La base de datos está compuesta por 9.800 extractos de vídeo con una duración de entre 8 y 12 segundos cada uno. Esta duración es suficiente para obtener extractos consistentes, lo cual permite que el espectador sienta emociones en este corto periodo de tiempo. Además, en estos fragmentos se ve reflejada la variedad de las películas que los componen, ya que contienen escenas violentas y sexuales, así como paisajes, entrevistas y otros tipos de escenas que reflejan la vida cotidiana.

Para justificar esta variedad de películas según su género se muestra en la siguiente ilustración una comparativa entre la distribución normalizada de películas según el género en LIRIS-ACCEDE frente a la misma distribución en las plataformas IMDB y ScreenRush (fuentes autorizadas más populares del mundo para el contenido cinematográfico y televisivo). Se puede observar la similitud en las distribuciones.

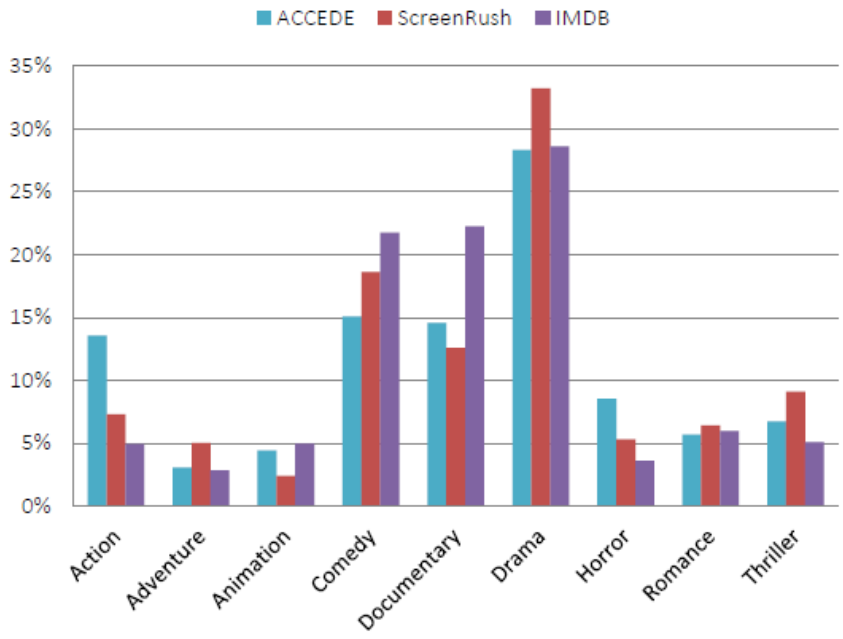


Ilustración 6. Comparativa en la distribución normalizada de películas por género entre LIRIS-ACCEDE, IMDB y ScreenRush [1]

Todos los extractos han sido etiquetados mediante crowdsourcing en la plataforma CrowdFlower [34]. La herramienta de evaluación empleada en esta tarea ha sido la de *Autoevaluación del maniquí* (comentada anteriormente en la *sección 3.1.3*). Como el concepto de *arousal* da pie a la ambigüedad, para asegurar la calidad de las etiquetas se empleó una herramienta avanzada de CrowdFlower denominada *"Quiz Mode":* los anotadores primero tenían que responder seis cuestiones tipo test y lograr una puntuación de un 70% o mayor con objeto de superar el cuestionario y poder así participar en la anotación. Esto asegura que sólo a los anotadores con altas calificaciones se les permite trabajar en la tarea.

Para la etiquetación del eje de la *valencia* participaron 1.517 anotadores de confianza pertenecientes a 89 países que obtuvieron una calificación en los test previos de 94.2% de precisión.

Sin embargo, como se ha mencionado anteriormente, la tarea de etiquetar el eje del *arousal* es algo más compleja y por ello participaron 2.442 anotadores de confianza también pertenecientes a 89 países.

Esta gran participación y diversidad cultural se refleja en la siguiente figura, donde se muestra el porcentaje de personas que participaron en el crowdsourcing según el país al que pertenecen.

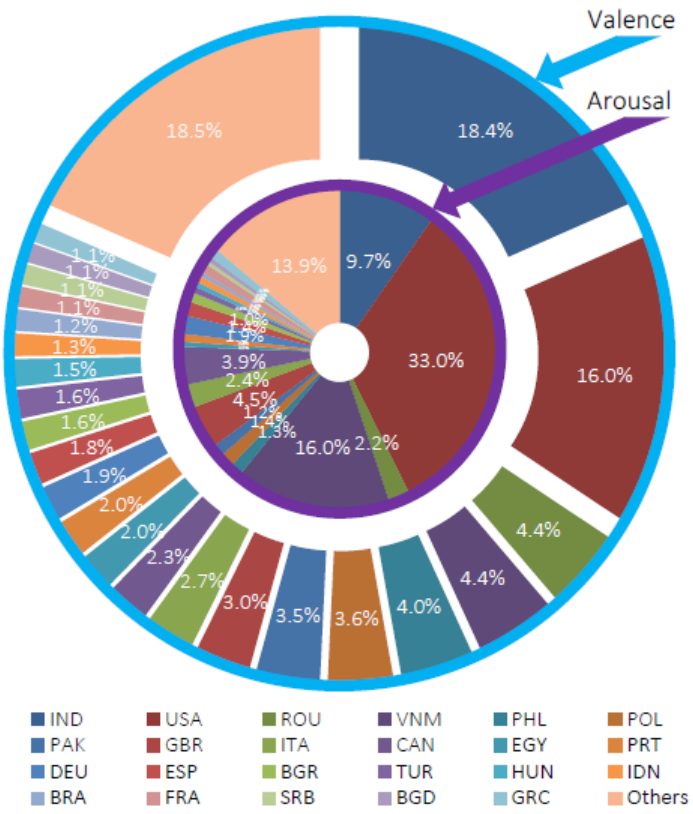


Ilustración 7. Países a los que pertenecen los anotadores de la *valencia* (círculo exterior) y del *arousal* (círculo interior) [1]

La combinación de la *valencia* y el *arousal* muestran resultados convincentes. Dietz y Lang mostraron en [35] que la *valencia* y el *arousal* están correlacionados y que las áreas de certeza de este espacio son más relevantes que las de otros.

En la siguiente ilustración se muestra el histograma en dos dimensiones (*valencia-arousal*) de los resultados obtenidos de las etiquetas. Cada celda indica el número de extractos de vídeo con una puntuación para *valencia* y para *arousal* entre los valores representados en ambos ejes. Hay que destacar que los valores mostrados en la *Ilustración 8* indican las posiciones relativas de los extractos en el espacio *valencia-arousal* y no su posición absoluta.

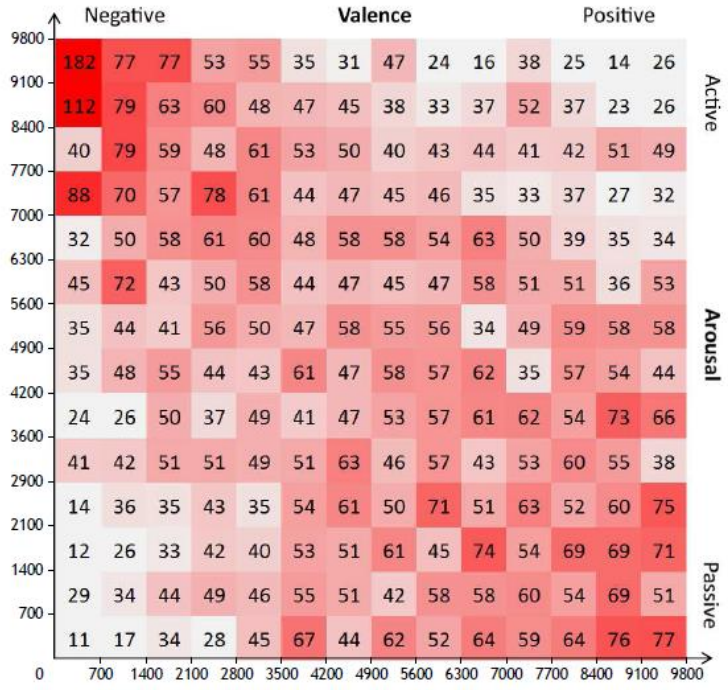


Ilustración 8. Histograma obtenido de los resultados de la etiquetación [1]

3.3.3 JUSTIFICACIÓN DE LA ELECCIÓN

La mayoría de las investigaciones que tratan modelos computacionales afectivos usan sus propias bases de datos, las cuales están diseñadas acorde a sus propios objetivos y necesidades. Este hecho hace que disminuya la eficiencia de la comunidad científica dedicada a los estudios afectivos. Según [1], lo que se busca con esta base de datos es que sea una referencia en un futuro y así los investigadores puedan disponer de una gran cantidad de datos etiquetados de forma certera con los que entrenar y probar sus aplicaciones.

Los motivos por los que se ha elegido esta base de datos son los siguientes:

- **Gran tamaño**. Gracias a los 9.800 extractos de vídeos que LIRIS-ACCEDE contiene hace que los estudios realizados sean más fiables al disponer de un gran *fondo de verdad* sobre el que entrenar y probar las investigaciones.

- **Libre distribución**. LIRIS-ACCEDE está compartida bajo las licencias de Creative Commons. Esto hace posible que sea de libre acceso y distribución y que no haya problemas de copyright. Por tanto toda la comunidad científica puede hacer uso de ella y los resultados pueden ser comparados entre investigadores que han trabajado sobre la misma base de datos.

- **Contenido diverso**. Los extractos de vídeo que forman la base de datos pertenecen a 160 películas, clasificadas según 9 géneros. El idioma principal de los vídeos es el inglés, pero también existe un conjunto con gran variedad de lenguas entre las que destacan francés, alemán, hindú, español e italiano. Toda esta variedad, tanto en el contenido como en el lenguaje, hace posible aumentar el *fondo de verdad* comentado anteriormente y por tanto mejorar la calidad de las investigaciones.

- **Etiquetada**. LIRIS-ACCEDE es una base de datos que está etiquetada fielmente en el espacio de dos dimensiones (2D) *valencia-arousal* por más de 4.000 personas mediante crowdsourcing. Se ha demostrado que las etiquetas que forman este espacio 2D están correlacionadas [35] y por tanto aportan mucha información a la hora de investigar sobre modelos computacionales de la emoción. Otro factor muy importante es la diversidad cultural que existe en las personas que han realizado la labor de etiquetado (ver *Ilustración 7*), ya que aporta fiabilidad en las etiquetas resultantes.

3.3.4 TRATAMIENTO DE LOS VÍDEOS

Previamente a la extracción de características acústicas de alto y bajo nivel de los vídeos es necesario analizarlos y separar el contenido de audio del contenido visual. Para ello se ha desarrollado un script en Matlab que va recorriendo el directorio en el que se encuentran los 9.800 vídeos que componen la base de datos LIRIS-ACCEDE, y haciendo uso de la herramienta MIRToolbox [2] se extrae el audio y los frames de cada vídeo. Los ficheros de audio resultantes se almacenan en otro directorio, mientras que los frames obtenidos son desechados al no resultar de interés en esta investigación.

De esta manera tenemos a nuestra disposición 9.800 archivos de audio (correspondientes a todos los vídeos de LIRIS-ACCEDE) sobre los que trabajaremos en este proyecto.

3.4 CARACTERÍSTICAS ACÚSTICAS DE BAJO NIVEL

Características acústicas de bajo nivel son todas aquellas características estadísticas que nos aporta la señal de audio tras haber superado un procesado básico de señal. Estas características son tales como la autocorrelación, desviación estándar, cruces por cero de la señal, energía, mfcc, factor de roll-off, brillo, etc.

Para la obtención de estas características se hace uso de la herramienta MIRtoolbox (la cual se describe en detalle a continuación), así como de los 9.800 ficheros de audio generados tras el análisis de la base de datos LIRIS-ACCEDE y el posterior tratamiento de sus vídeos.

3.4.1 MIRTOOLBOX

MIRtoolbox [2] es una herramienta para Matlab dedicada a la extracción de características acústicas procedentes de archivos de audio. Ha sido diseñado específicamente con el objetivo de permitir la computación de un gran rango de características de bases de datos de archivos de audio, que puedan ser aplicadas a análisis estadísticos.

La principal ventaja que ofrece MIRtoolbox con respecto a otras herramientas similares es que está desarrollado para ser usado en Matlab. Esto hace que otros *toolbox* pertenecientes a MathWorks, como *Statistics toolbox* o *Neural Network toolbox*, puedan ser usados para analizar en profundidad las características extraídas mediante MIRtoolbox sin necesidad de tener que exportar datos de un software a otro.

Otra ventaja frente a otros software es los procesos analíticos complejos pueden ser diseñados con una sintaxis muy simple, cuyo poder expresivo viene del uso de un paradigma orientado a objetos.

Las diferentes características musicales extraídas de los archivos de audio son altamente interdependientes: en particular, como se puede apreciar en la *Ilustración 9,* algunas características están basadas en las mismas computaciones iniciales. Con el objeto de mejorar la eficiencia computacional, es importante evitar computación redundante en estos componentes comunes. Cada uno de estos componentes intermedios, y las características musicales finales, son considerados como *building blocks* que pueden ser libremente combinados unos con otros. Por otro lado, manteniendo el objetivo de optimizar la facilidad de uso de la herramienta, cada *building block* ha sido concebido de tal forma que es capaz de adaptarse al tipo de datos de entrada. Por ejemplo, la computación de los MFCCs puede estar basada en la forma de onda de la señal de audio inicial, o en representaciones intermedias como el espectro o el *mel-scale* espectro (ver *Ilustración 9*). De forma similar, la autocorrelación está calculada a través de rangos diferentes de *delays* dependiendo del tipo

de datos de entrada (forma de onda del audio, espectro, envolvente de la señal). Esta descomposición de todo el conjunto de algoritmos de extracción de características en un conjunto común de *building blocks* tiene la ventaja de ofrecer un resumen sintético de los diferentes alcances estudiados en este dominio de la investigación.

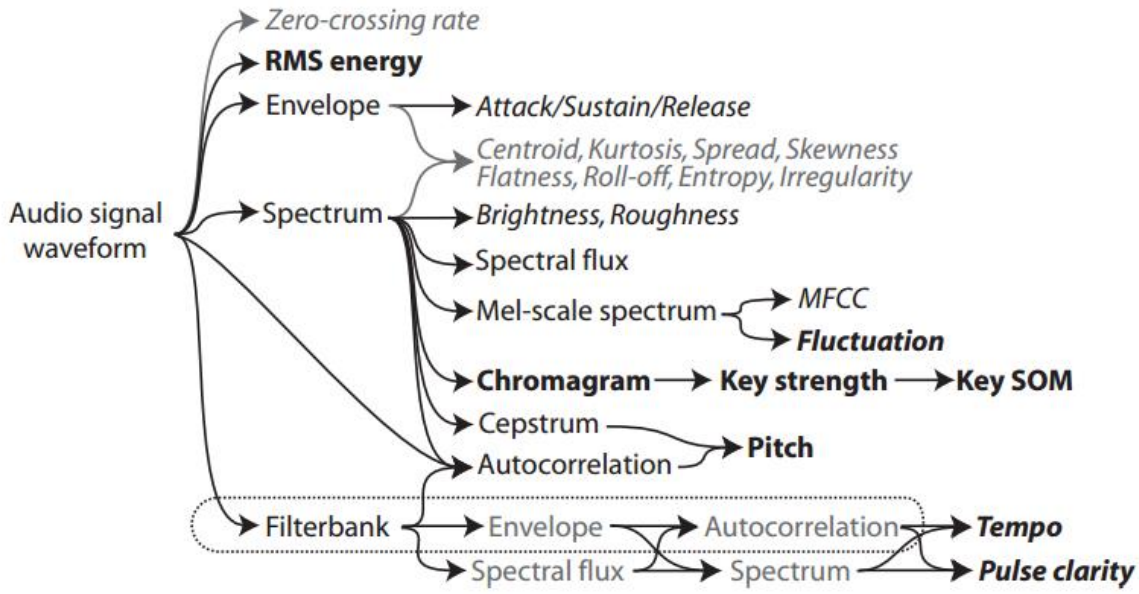


Ilustración 9. Resumen de las características musicales que pueden ser extraídas mediante MIRtoolbox [2]

En esta ilustración vemos un resumen de las características principales implementadas en la herramienta. Todos los procesos comienzan con la señal de audio (izquierda) y van formando una cadena de operaciones hacia la derecha. La disposición vertical de los procesos indica el orden de complejidad de las operaciones, siendo la computación más simple la que está más arriba y la más compleja la que está en la parte inferior.

Cada característica musical está relacionada a una de las dimensiones musicales tradicionales definidas en la teoría musical. Las palabras escritas en negrita indican características relacionadas con la tonalidad (*Chromagran*, *Key strength* y *Key SOM*) y con la dinámica (*RMS energy*). Las palabras escritas en negrita y además en cursiva indican características relacionadas con el ritmo, con el tempo, con la claridad del pulso y con la fluctuación. Las palabras escritas únicamente en cursiva remarcan un gran conjunto de características que pueden ser asociadas al timbre. Entre ellas, todos los operadores indicados mediante palabras escritas en cursiva y de color gris pueden ser aplicados a muchas otras representaciones diferentes: por ejemplo, es posible aplicar aspectos estadísticos como el centroide a cualquier espectro, o envolvente, pero también a histogramas basados en cualquier característica dada.

En el apartado *Apéndices* se muestra un esquema con todas las características contenidas en la herramienta MIRtoolbox.

3.4.2 EXTRACCIÓN DE CARACTERÍSTICAS

Una vez comprendido el uso y visto el potencial del que dispone MIRtoolbox, se empleará esta herramienta para calcular 392 características acústicas por cada fichero de audio que hemos obtenido como resultado de la separación de la parte acústica y visual del vídeo (*sección 3.3.4*).

Este conjunto de características acústicas obtenidas se denominarán de ahora en adelante como *características acústicas de bajo nivel*.

El resultado de este proceso de extracción de características da como resultado 9.800 ficheros con extensión "arff" [38] que serán almacenados en un nuevo directorio para su posterior uso. Tras esto, se realiza un script en Python que permite disponer de un solo archivo "arff" con todas las características acústicas de bajo nivel.

3.5 CARACTERÍSTICAS ACÚSTICAS DE ALTO NIVEL

Definiremos características acústicas de alto nivel como aquellas características acústicas que pueden ser propuestas una vez que el fichero de audio haya superado una etapa de segmentación.

El objetivo de esta etapa de segmentación es conocer los diferentes eventos acústicos que componen el archivo de audio y la duración de cada evento dentro del mismo archivo. Los eventos acústicos que diferenciaremos serán *speech*, *music*, *speech-music*, *others*.

Tras tener todos los audios pertenecientes a los vídeos de la base de datos LIRIS-ACCEDE [1] segmentados, se propondrán características acústicas de alto nivel y se realizará una experimentación para comprobar si estas características propuestas influyen en la respuesta afectiva de un espectador al visualizar un vídeo o por el contrario pueden ser descartadas.

Algunos ejemplos de características acústicas de alto nivel que pueden ser propuestas son: distinción entre una persona hablando y música, porcentaje de un evento acústico concreto dentro de un fichero de audio, o apreciación de varios interlocutores en un fragmento de solo voz.

La herramienta de segmentación que se ha empleado es la misma que A. Gallardo and R. San Segundo presentaron al concurso de segmentación de audio *Albayzín evaluation 2010*

on Audio Segmentation, en el año 2010 [8] [9]. En dicho concurso, esta herramienta obtuvo la mejor puntuación con un 30.22% de tasa media de error.

La herramienta usada para la evaluación de la segmentación será una herramienta similar a la empleada en el concurso de Albayzín [9]. Con ella seremos capaces de comprobar qué error obtenemos en la segmentación y qué evento acústico es el que más falla o el que mejor funciona.

Lo primero que encontraremos en esta sección es un estudio de la herramienta de segmentación [8] y las consideraciones que se tendrán en cuenta. Después comprobaremos los resultados obtenidos en la segmentación mediante la herramienta empleada en [9] y estudiaremos si los resultados obtenidos son aceptables para continuar con la investigación. Por último se propondrán características acústicas de alto nivel y se integrarán en un fichero "arff" para su futura experimentación.

3.5.1 SEGMENTACIÓN

La herramienta de segmentación empleada surge como propuesta al concurso *Albayzín evaluation 2010 on Audio Segmentation* [9]. A. Gallardo and R. San Segundo presentaron a concurso la herramienta [8], con la cual obtuvieron el primer puesto.

El sistema propuesto está basado en Modelos Ocultos de Markov (HMMs), incluyendo un HMM de 3 estados por cada clase acústica. Para la extracción de características, se consideraron términos estadísticos a largo plazo de MFCC (Mel Frequency Ceptral Coefficients), entropía espectral y coeficientes de croma. Las características del croma son una poderosa representación para estudios acústicos, ya que el espectro es dividido en 12 partes que representan 12 semitonos distintos (o croma) de la octava musical.

La tarea de evaluación propuesta en [9] consiste en la segmentación de audios pertenecientes a noticias de radio en diferentes clases acústicas (ACs):

- Speech [sp]. Voz clara en estudio captada con un micrófono situado cerca del interlocutor.
- Music [mu]. Música entendida en el caso general.
- Speech with noise in background [sn]. Voz que no está grabada en condiciones de estudio, o está superpuesta con algún tipo de ruido (aplausos, ruido de tráfico, etc.), o incluye varias voces simultáneamente (por ejemplo, una traducción en tiempo real).
- Speech with music in background [sm]. Superposición de las clases *speech* y *music* o de las clases *speech with noise in background* y *music*.

Existe también otra clase acústica que no es evaluada: Other [ot]. Esta clase se refiere a cualquier tipo de señal de audio (incluido ruido) que no se corresponde a ninguna otra clase.

La base de datos usada para desarrollar y entrenar el sistema de segmentación que A. Gallardo and R. San Segundo propusieron en [8] consiste en una base de datos de noticias catalanas pertenecientes al canal de televisión 3/24 que fueron grabadas por el Centro de Investigación TALP de la UPC, y fueron anotadas por Verbio Technologies [10]. La Corporació Catalana de Mitjans Audiovisuals, propietaria del contenido multimedia, permite su uso para investigación y desarrollo tecnológico. La base de datos, que incluye alrededor de 87 horas de sonido (24 archivos de aproximadamente 4 horas de duración), fue dividida en dos partes: para entrenamiento/desarrollo (2/3 del total de los datos disponibles), y para pruebas (el restante 1/3). El idioma de los archivos de audio se corresponde con un 83% a catalán y con un 17% a español. La distribución de las clases acústicas en la base de datos es la siguiente: 37% *speech*, 5% *music*, 15% *speech with music in background*, 40% *speech with noise in background*, 3% *other*. Las señales de audio están en formato pcm, mono, 16 bit de resolución, y muestreada a una frecuencia de 16 kHz.

La herramienta de segmentación [8] está conformada por un sistema *one-step* basado en HMM. En particular, se ha considerado un HMM de 3 estados por cada clase acústica, considerando 16 Gaussianas por estado. La topología del HMM se puede ver en la siguiente ilustración (ver *Ilustración 10*).

Las características consideradas en este sistema han sido procesos estadísticos aplicados con una ventana de 1 segundo de duración (con una superposición de 0.5 segundos) sobre 15 MFCCS (Mel Frequency Cepstral Coefficients) y la energía local calculada en ventanas de 25 milisegundos (con una superposición de 15 milisegundos), y su delta y doble delta. Los procesos estadísticos aplicados son la media y la desviación estándar. En total, hay 96 características cada 0.5 segundos.

Para todos los experimentos, se ha usado el software HTK [23] para entrenar y probar los HMMs. Para extracción de características, se ha empleado la herramienta OpenSMILE [22].

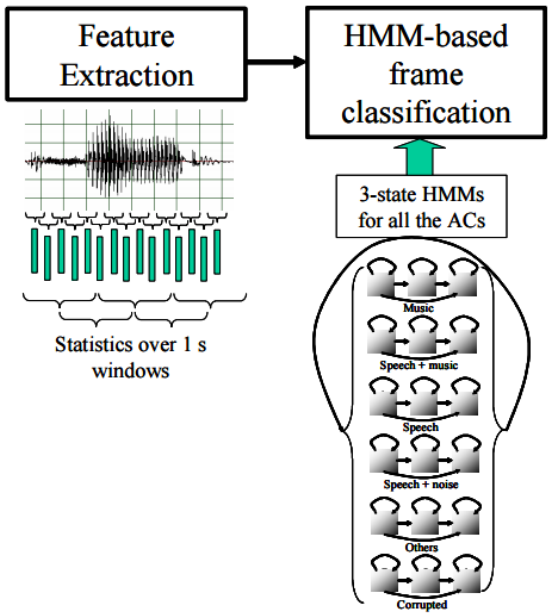


Ilustración 10. Diagrama del sistema con detalles sobre la extracción de características y topología de los HMM [8]

Como se ha mencionado anteriormente, en este proyecto se hará uso de la herramienta de segmentación propuesta en [8].

El proceso se describe a continuación:

Tras haber separado los frames del audio de cada vídeo en la *sección 3.3.4* y disponer en un directorio los 9.800 ficheros de audio correspondientes a todos los videos de LIRIS-ACCEDE, el primer paso es submuestrear todas las señales de audio disponibles a 16 kHz para que puedan ser tratadas por el sistema de segmentación. Para ello se emplea el software SoX [26], definido por sus creadores como "La navaja suiza de los programas de procesamiento de sonido". Esta herramienta que funciona a través de la línea de comandos es capaz de convertir archivos de audio a varios formatos, así como aplicar efectos a esos archivos, y reproducir y grabar ficheros de audio en la mayoría de las plataformas.

Una vez disponibles los 9.800 ficheros de audio submuestreados a 16 kHz es necesario parametrizarlos para obtener las características estadísticas de la señal que compone cada uno. Siguiendo la línea de [8] para desarrollar la herramienta de segmentación haremos uso del software OpenSMILE [22], que es una herramienta que permite extraer gran cantidad de características de audio en tiempo real. Combina características de recuperación de información musical (*Music Information Retrieval*) y de procesado de voz (*Speech Processing*). Al finalizar esta parametrización dispondremos de un directorio con 9.800 ficheros de formato "par".

Por último, se empleará la herramienta HTK [23] con los parámetros de configuración del estudio [8] para segmentar todos los ficheros de audio parametrizados en formato "par"

obtenidos anteriormente. HTK consiste en un conjunto de módulos de librerías y herramientas que proporcionan sofisticadas facilidades para el análisis de la voz, el entrenamiento de Modelos Ocultos de Markov (HMM), *testing* y análisis de los resultados. El software soporta HMMs usando tanto la mezcla de densidad continua de Gaussianas como distribuciones discretas y puede ser usado para construir complejos sistemas HMM.

A la hora de realizar este último paso, hay que introducir en la línea de configuración de HTK dos pesos, los cuales se corresponden al peso de inserción y al modelo de lenguaje. La variación de estos pesos influye en el resultado final de la segmentación de forma que si disminuimos su valor, la herramienta segmentará en más fragmentos el fichero de audio, y si aumentamos su valor, el resultado de la segmentación será de menos fragmentos por fichero de audio. Es por tanto necesario ajustar los valores de estos pesos para encontrar el resultado de la segmentación más óptimo para el desarrollo de nuestro proyecto, para ello se seleccionarán los pesos que hacen mínimo el ratio medio de error en la evaluación de la segmentación.

El resultado de esta segmentación es un archivo de tipo "mlf" en el que se encuentran todos los ficheros de audio segmentados conforme las clases acústicas comentadas anteriormente: *sp*, *mu*, *sm*, *sn*.

Tras la obtención del fichero "mlf", se realizará un script en Python cuya función será parsear los datos contenidos en el archivo "mlf" al formato requerido por la herramienta de la evaluación de segmentación.

En el diagrama mostrado a continuación se refleja el desarrollo de la segmentación descrito anteriormente:

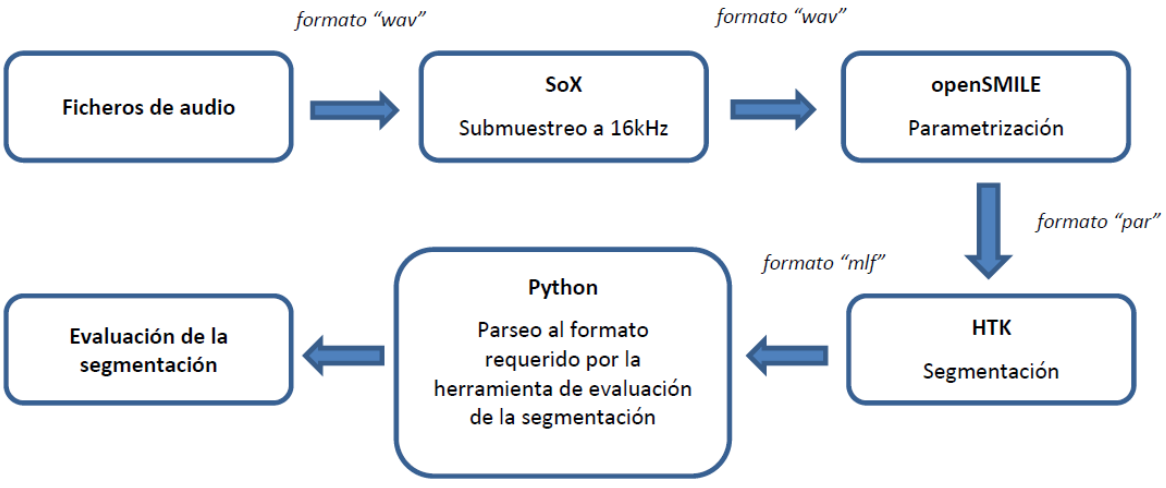


Ilustración 11. Diagrama de la segmentación de un fichero de audio

En la siguiente ilustración podemos ver un ejemplo del resultado de la segmentación (archivo "mlf"), en la que solo se muestra una pequeña parte de los 9.800 audios segmentados. La primera línea que se muestra es el nombre del fichero de audio, las siguientes líneas corresponden a las distintas clases acústicas que componen cada fichero de audio. El primer número corresponde al tiempo de inicio del fragmento, el segundo número corresponde al tiempo final del fragmento (para obtener el tiempo en segundos es necesario dividir el número entre diez elevado a la menos 7), la tercera anotación se corresponde con la clase acústica correspondiente al fragmento, y el último número es el *score* o puntuación. Este *score* se refiere a la "confianza" con la que los HMM empleados en el sistema de segmentación han decidido que el evento acústico es el correspondiente y no otro.

En el caso concreto de *"tmp_ACCEDE00010.wav.rec"* se aprecia que el fichero de audio ha sido dividido en dos fragmentos porque aparecen dos eventos acústicos distintos. Por el contrario se aprecia que *"tmp_ACCEDE00011.wav.rec"* está compuesto por solo un fragmento.

```
"tmp_ACCEDE00009.wav.rec"
0 90000000 others 316.199615
.
"tmp_ACCEDE00010.wav.rec"
0 20000000 music 79.446098
20000000 100000000 others 410.034058
.
"tmp_ACCEDE00011.wav.rec"
0 105000000 speechmusic -174.248093
.
"tmp_ACCEDE00012.wav.rec"
0 10000000 speech -26.730404
10000000 105000000 others 1055.848022
.
"tmp_ACCEDE00013.wav.rec"
0 75000000 music 310.267975
.
"tmp_ACCEDE00014.wav.rec"
0 95000000 music -265.557495
```

Ilustración 12. Ejemplo de archivo "mlf", resultante de la segmentación

Para continuar con el desarrollo del trabajo y sin distraernos de nuestro objetivo (proponer características acústicas de alto nivel), es fundamental realizar una evaluación de la segmentación realizada para saber si se ha hecho de forma correcta o por el contrario el ratio medio de error es demasiado elevado. Se evaluará a continuación.

3.5.2 EVALUACIÓN DE LA SEGMENTACIÓN

La herramienta para evaluar el proceso de segmentación llevado a cabo en la anterior sección será similar a la empleada en el concurso *Albayzín evaluation 2010 on Audio Segmentation* [9]. La medida es definida como el error relativo medio sobre todas las clases acústicas:

$$Error = \underset{i}{average}\left(\frac{dur(miss_i) + dur(fa_i)}{dur(ref_i)}\right)$$

Ilustración 13. Fórmula para calcular el error relativo medio

Donde:

- *dur(miss$_i$)* es la duración total de todos los errores correspondientes a pérdidas para la *iésima* clase acústica.
- *dur(fa$_i$)* es la duración total de todos los errores correspondientes a falsas alarmas para la *iésima* clase acústica.
- *dur(ref$_i$)* es la duración total de todas las *iésimas* instancias de clase acústica acorde al fichero de referencia.

La incorrecta clasificación de un segmento de audio es calculada tanto como un error de pérdida en una clase acústica y como un error de falsa alarma para otra. Existe además un parámetro denominado *collar*, de 1 segundo de duración. Quiere decir que el primer segundo del fragmento y el último segundo del fragmento no se tendrán en cuenta a la hora de evaluar el error. Esto se hace para minimizar el error que pueda surgir debido a la inconsistencia de la anotación humana y a la no certeza de cuándo una clase acústica comienza o termina.

Esta herramienta ha sido desarrollada en Matlab por T. Butko para la evaluación del concurso de segmentación [9]. Para poder hacer uso de ella es necesario disponer de dos archivos de entrada, cuyos formatos deben de ser:

nombre_fichero tiempo_inicio tiempo_final etiqueta

Ha sido necesario desarrollar un script en Python para la realización de esta tarea. Como resultado de la segmentación habíamos obtenido un fichero en formato “mlf” con 9.800 instancias, como el mostrado en la *Ilustración 12*. Tras la programación de la tarea en Python se consigue un resultado como el mostrado en la siguiente figura, el cual es el correcto para introducir el fichero en la herramienta de evaluación de segmentación.

```
ACCEDE01000 0.0 7.5 sn
ACCEDE01100 0.0 8.5 sm
ACCEDE01200 0.0 7.5 sp
ACCEDE01300 0.0 11.5 sn
ACCEDE01400 0.0 3.5 sm
ACCEDE01400 3.5 9.0 sn
ACCEDE01500 0.0 2.0 mu
ACCEDE01500 2.0 8.0 sm
ACCEDE01600 0.0 8.5 sm
ACCEDE01700 0.0 10.0 sn
```

Ilustración 14. Formato necesario para introducir los ficheros en la herramienta de evaluación de la segmentación

Este archivo es uno de los dos que hay que introducir en la función desarrollada en Matlab para evaluar la segmentación. Se llamará *archivo hipotético*, ya que es el que obtenemos tras realizar la segmentación.

El otro fichero que hay que introducir en la herramienta es el denominado *archivo referencia*. Este archivo deber ser el resultante a una segmentación realizada de forma manual, anotada por una persona. Así podremos saber el error medio que ha sufrido la segmentación realizada por la herramienta en comparación a la segmentación que haría una persona. Para crear el archivo de referencia se han etiquetado y segmentado de forma manual una selección de los 9.800 vídeos contenidos en la base de datos LIRIS-ACCEDE, en concreto 100 vídeos. Estos 100 vídeos se han escogido espaciados entre sí para que haya diversidad en su contenido al no pertenecer dos fragmentos de los seleccionados a la misma película. Esta elección asegura también diversidad en el idioma, así como en los eventos acústicos que pueden formar los fragmentos, siendo más fiable al haber más variedad. En cuanto a los anotadores (personas que realizan la etiquetación manualmente), los 100 vídeos han sido anotados por el autor de este trabajo. Lo ideal es que cuantos más anotadores haya y más vídeos se etiqueten, de más calidad serán los datos sobre los que se trabajen, pero los recursos son limitados y no ha sido posible encontrar más anotadores.

3.5.3 RESULTADOS

En primer lugar, vamos a mostrar los resultados que obtuvo la herramienta de segmentación [8] (que es la que ha sido usada en este proyecto) en el concurso de evaluación de segmentación [9] (evaluado con la misma herramienta que emplearemos sobre nuestra segmentación realizada). Los resultados son los que se muestran a continuación, y serán denominados como *Evaluación 1*:

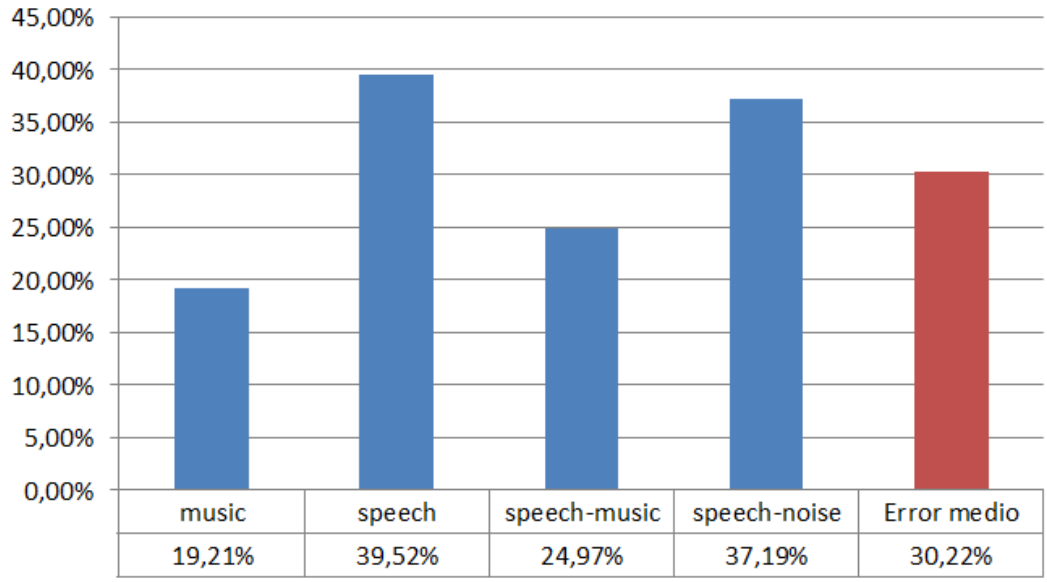


Ilustración 15. Evaluación 1: Resultados de la herramienta de segmentación [8] en el concurso de evaluación [9]

Se obtuvo un error medio de 30.22%, habiendo una clara diferencia entre las clases acústicas que contienen música y las que contienen voz (aproximadamente un 20% de error de diferencia).

Ahora, habiendo ajustado los pesos de segmentación para nuestro sistema, procedemos a evaluar la calidad de nuestra segmentación. La siguiente gráfica muestra los resultados, que serán denominados *Evaluación 2*:

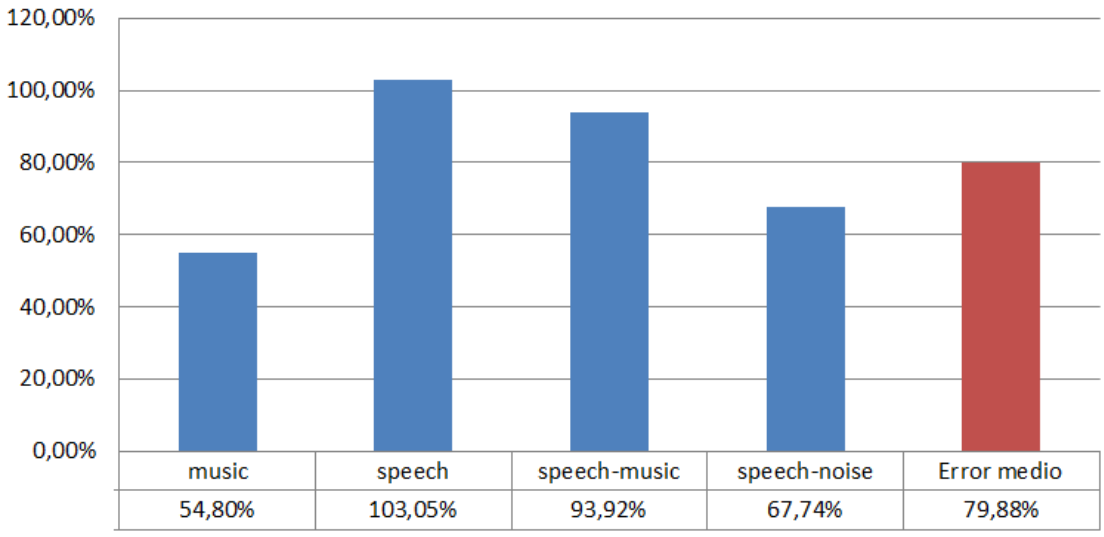


Ilustración 16. Evaluación 2: Resultados de la evaluación de la segmentación para el sistema propuesto en este proyecto

Podemos observar que el error medio obtenido es de casi un 80%, valor que supera en aproximadamente un 50% el error obtenido por [8] en la evaluación [9]. Esto nos indica que la segmentación realizada no tiene suficiente validez como para extraer de estos datos obtenidos características acústicas de alto nivel.

Como consecuencia de este resultado se realiza una investigación exhaustiva de la herramienta de segmentación empleada para intentar aclarar qué puede suceder y cómo es posible mejorar el ratio medio de error.

Lo primero que hay que destacar es que la herramienta de segmentación ha sido diseñada para segmentar audios pertenecientes a la base de datos TALP [10], en los que un 83% de ellos son en catalán y el resto en español, y por tanto sus modelos (HMMs) han sido entrenados bajo estas condiciones.

Nosotros estamos empleando la misma herramienta (con los modelos entrenados para segmentar noticias principalmente en catalán) para la segmentación de los audios pertenecientes a la base de datos LIRIS-ACCEDE. Como ya se ha comentado anteriormente (ver *sección 3.3.2*) esta base de datos es de contenido muy diverso, en la que hay películas distinguidas en 9 géneros y cuyo idioma principal es el inglés, habiendo pinceladas de otros idiomas (español, italiano, francés, etc.). Otro apunte que cabe destacar es que LIRIS-ACCEDE contiene fragmentos de películas, en las que en la mayoría de las veces se busca conseguir algún efecto o cierta expectación del espectador y para ello se crean efectos sonoros muy diferentes a los que encontramos en los audios de un noticiero (como los pertenecientes a la base de datos TALP).

Además, analizando el archivo de referencia introducido en la herramienta que evalúa la segmentación frente al archivo hipotético introducido se ha visto que la mayoría de los errores proceden de la clase acústica *speech-noise.*

Con la suma de todas estas premisas, se realizan tres ajustes o modificaciones con la intención de mejorar el ratio de error medio de la segmentación.

El primero de ellos consiste en cambiar todas las clases acústicas de tipo *speech-noise* por tipo *speech*. Este ajuste se realiza tanto sobre el archivo de referencia como sobre el hipotético que vamos a evaluar.

El segundo consiste en añadir la clase acústica *others* y asemejarla a cualquier evento acústico que no pueda ser etiquetado como ninguno de los demás existentes. Esta clase es muy útil porque aporta todos los eventos acústicos que son solo ruidos, o momentos en silencio en los vídeos que encontramos en LIRIS-ACCEDE y que son muy difíciles de encontrar en la base de datos TALP (al ser un noticiero). Esta modificación también se aplica al archivo de referencia y al archivo hipotético que se va a evaluar.

La tercera y última modificación se realiza sobre el *collar* de tiempo de 1 segundo de duración que permitía evitar a la hora de hacer la evaluación de la segmentación el primer y el último segundo de cada fragmento de audio con el fin de minimizar el error. Se aumenta el valor de tiempo del *collar* de 1 segundo a 2.5 segundos para mejorar la inconsistencia de la anotación del fichero de referencia, ya que éste ha sido anotado únicamente por una persona.

Con estas modificaciones llevadas a cabo, se obtienen los siguientes resultados denominados *Evaluación 3*:

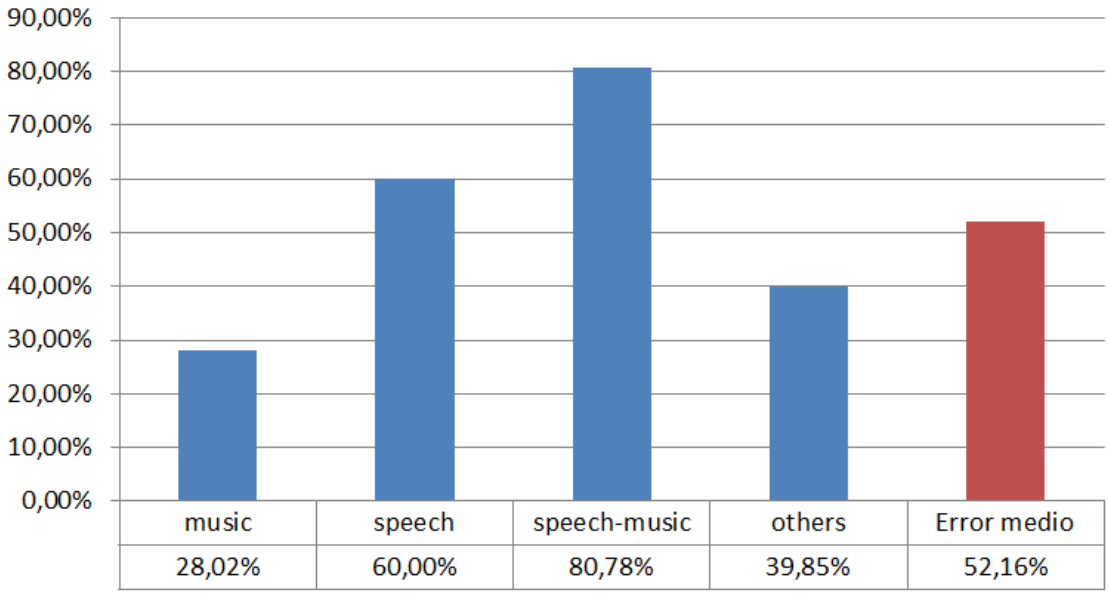
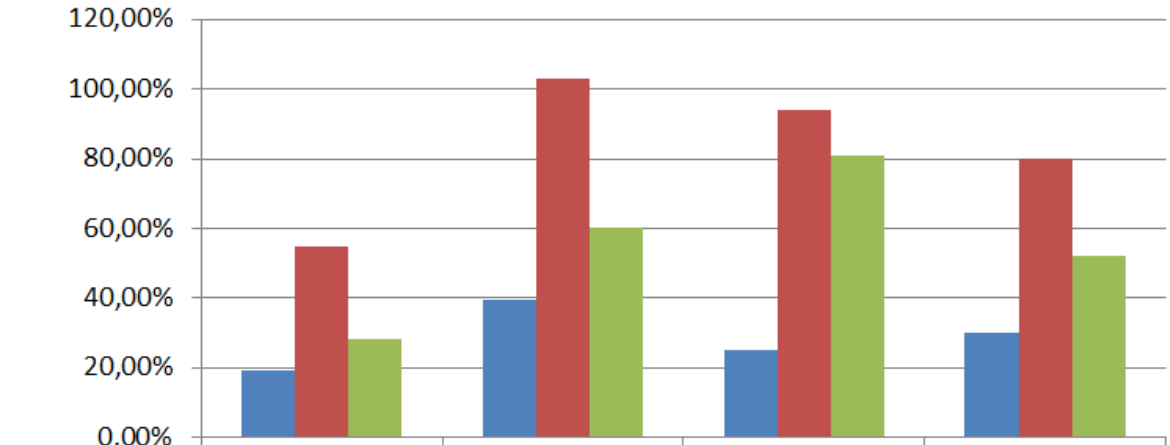


Ilustración 17. Evaluación 3: Resultados de la evaluación de la segmentación con la aplicación de 3 modificaciones

Se observa que el error medio ha disminuido aproximadamente un 30% con respecto al caso anterior. Esto indica que los ajustes realizados funcionan correctamente y mejoran la segmentación.

A continuación, se realiza una comparativa de las tres evaluaciones llevadas a cabo:



	music	speech	speech-music	Error medio
Evaluación 1	19,21%	39,52%	24,97%	30,22%
Evaluación 2	54,80%	103,05%	93,92%	79,88%
Evaluación 3	28,02%	60,00%	80,78%	52,16%

Ilustración 18. Comparativa de las evaluaciones de la segmentación llevadas a cabo

La clase acústica *speech-noise* ha sido suprimida debido a que en la evaluación 1 y en la 2 aparece, pero después de hacer la asimilación de *speech-noise* a *speech* y añadir la clase *others* en la evaluación 3 ya no aparece.

Se aprecia en la comparativa que tanto en la *Evaluación 1* como en la *Evaluación 3*, el error obtenido en las clases acústicas *speech* y *music* asciende de manera proporcional. Además, tanto para nuestros resultados como para los obtenidos en la *Evaluación 1* el evento acústico con menos ratio de error es el *speech*, lo que significa que nuestra herramienta es capaz de diferenciar razonablemente bien un fragmento que contiene voz de otro que no tiene voz. Sin embargo, el máximo error obtenido se da en la clase *speech-music*, lo cual nos hace pensar en la composición de la base de datos LIRIS-ACCEDE, donde al ser películas, los directores "juegan" con la combinación de música y voz para crear un tipo de sentimiento o afección en el espectador. Conociendo esto, podemos decir que es normal que esta clase sea la más problemática y que haga aumentar el error medio.

En este mal resultado de la clase acústica *speech-music* también influye que estamos usando una herramienta de segmentación, cuyos modelos han sido entrenados para segmentar audios en catalán, para segmentar audios en los que aparecen gran variedad de idiomas (alemán, francés, italiano, español, etc.).

Con estas premisas podemos decir que la herramienta es capaz de segmentar de forma certera entre diferentes clases acústicas, sobretodo diferencia muy bien los fragmentos en los que aparece voz de los que aparece música. El problema es que al estar entrenado el sistema para segmentar otro tipo de audios, carece de carácter suficiente para etiquetar

correctamente a qué clase acústica pertenece cada fragmento. Este problema también hace que el error medio aumente considerablemente.

También podemos observar que el error medio de la *Evaluación 3* (52.16%) está más cercano al error medio de la *Evaluación 1* (30.22%) que al de la *Evaluación 2* (79.88%).

Tras este estudio en profundidad de los resultados obtenidos y teniendo en cuenta todas las consideraciones mencionadas anteriormente, podemos afirmar que el resultado obtenido tras realizar los ajustes es razonable para continuar con la investigación y proponer características acústicas de alto nivel basadas en la segmentación realizada.

3.5.4 PROPUESTA DE CARACTERÍSTICAS ACÚSTICAS DE ALTO NIVEL

En esta sección se propondrán características acústicas de alto nivel derivadas de la segmentación de los ficheros de audio realizada anteriormente. Como la evaluación de la segmentación ha sido razonable, podremos proponer características relacionadas con las clases acústicas definidas en nuestro sistema (*speech*, *music*, *speech-music*, *others*) y más adelante experimentar con ellas para comprobar si éstas influyen en la respuesta afectiva de un espectador que visualiza un vídeo o por el contrario no aportan información interesante que pueda ser estudiada.

Las características acústicas de alto nivel propuestas son 19. Se definen a continuación:

- **others**. De valor binario. Será 1 si el fragmento pertenece a la categoría acústica *others* y 0 sino pertenece a dicha categoría.
- **speechmusic.** De valor binario. Será 1 si el fragmento pertenece a la categoría acústica *speech-music* y 0 sino pertenece a dicha categoría.
- **music.** De valor binario. Será 1 si el fragmento pertenece a la categoría acústica *music* y 0 sino pertenece a dicha categoría.
- **speech.** De valor binario. Será 1 si el fragmento pertenece a la categoría acústica *speech* y 0 sino pertenece a dicha categoría.
- **more_than_one_fragments.** De valor binario. Será 1 si el fichero de audio está compuesto por 2 o más fragmentos.
- **percentaje_others.** De valor numérico. Es el porcentaje correspondiente a la clase acústica *others* con respecto a la duración total del fragmento de audio.
- **percentaje_speechmusic.** De valor numérico. Es el porcentaje correspondiente a la clase acústica *speech-music* con respecto a la duración total del fragmento de audio.
- **percentaje_music.** De valor numérico. Es el porcentaje correspondiente a la clase acústica *music* con respecto a la duración total del fragmento de audio.
- **percentaje_speech.** De valor numérico. Es el porcentaje correspondiente a la clase acústica *speech* con respecto a la duración total del fragmento de audio.
- **number_of_fragments.** De valor numérico. Corresponde al número de fragmentos que componen el archivo de audio segmentado.
- **score_others.** De valor numérico. Corresponde al valor numérico de la "confianza" que tiene el sistema para decidir la clase acústica *others* con respecto a la duración total del fragmento de audio.
- **score_speechmusic.** De valor numérico. Corresponde al valor numérico de la "confianza" que tiene el sistema para decidir la clase acústica *speech-music* con respecto a la duración total del fragmento de audio.
- **score_music.** De valor numérico. Corresponde al valor numérico de la "confianza" que tiene el sistema para decidir la clase acústica *music* con respecto a la duración total del fragmento de audio.

- **score_speech.** De valor numérico. Corresponde al valor numérico de la "confianza" que tiene el sistema para decidir la clase acústica *speech* con respecto a la duración total del fragmento de audio.
- **score_global.** De valor numérico. Corresponde al valor numérico total de la "confianza" que tiene el sistema a la hora de decidir todos los eventos acústicos que componen el archivo de audio.
- **only_others.** De valor binario. Será 1 si en el fichero de audio sólo aparece la clase acústica *others*.
- **only_speechmusic.** De valor binario. Será 1 si en el fichero de audio sólo aparece la clase acústica *speech-music*.
- **only_music.** De valor binario. Será 1 si en el fichero de audio sólo aparece la clase acústica *music*.
- **only_speech.** De valor binario. Será 1 si en el fichero de audio sólo aparece la clase acústica *speech*.

Para el posterior estudio de estas características se desarrolla un script en Python que nos da como resultado un archivo con formato "arff". Este formato, explicado en detalle en la *sección 4.2.1*, es el requerido por la herramienta de experimentación WEKA, la cual usaremos para el proceso de experimentación más adelante. El archivo "arff" resultante del script creado tendrá como atributos los nombres de las características acústicas de alto nivel descritas anteriormente y como valores los correspondientes al fichero que menos error medio ha conseguido de la evaluación de la segmentación.

CAPÍTULO 4. EXPERIMENTACIÓN. APRENDIZAJE MÁQUINA: WEKA

Tras haber extraído características acústicas de bajo nivel y haber propuesto y extraído características acústicas de alto nivel, el siguiente paso consiste en realizar una serie de experimentos con los que podamos conocer si tanto las características acústicas de bajo nivel como las propuestas de alto nivel influyen en la respuesta afectiva de un espectador al visualizar un vídeo. Éste es el objetivo principal del proyecto, ya que comprobaremos si todo el trabajo desarrollado anteriormente supone una contribución en la investigación de este campo o no.

Primero se realiza una introducción sobre clasificación y aprendizaje máquina para conocer cómo funciona el sistema de entrenamiento, clasificación y experimentación. Después se presentará el software empleado, WEKA [25], con el que se ha trabajado en esta parte de la investigación. En el siguiente apartado se explicará cómo se han obtenido las etiquetas de la base de datos LIRIS-ACCEDE [1] para realizar la clasificación. Por último se mostrarán los experimentos que se han llevado a cabo así como su resultado.

4.1 CLASIFICACIÓN

Es un proceso que se incluye dentro del aprendizaje máquina. El aprendizaje máquina es un campo de la inteligencia artificial que estudia algoritmos capaces de aprender de datos de entrada, extrayendo ciertas características de ellos con el objetivo de hacer predicciones sobre nuevos datos de entrada. Según Arthur Samuel, pionero en el campo de la inteligencia artificial, “el aprendizaje máquina proporciona a los ordenadores la capacidad de aprender por sí mismos sin haber sido programados”.

Este proceso está compuesto por dos etapas diferentes: lo primero, el clasificador necesita ser entrenado, proporcionándole un conjunto de datos de entrada. Después de esto, el clasificador va a ser probado usando el modelo entrenado previamente e introduciéndole un nuevo conjunto de datos. El proceso se explica con mayor detalle a continuación.

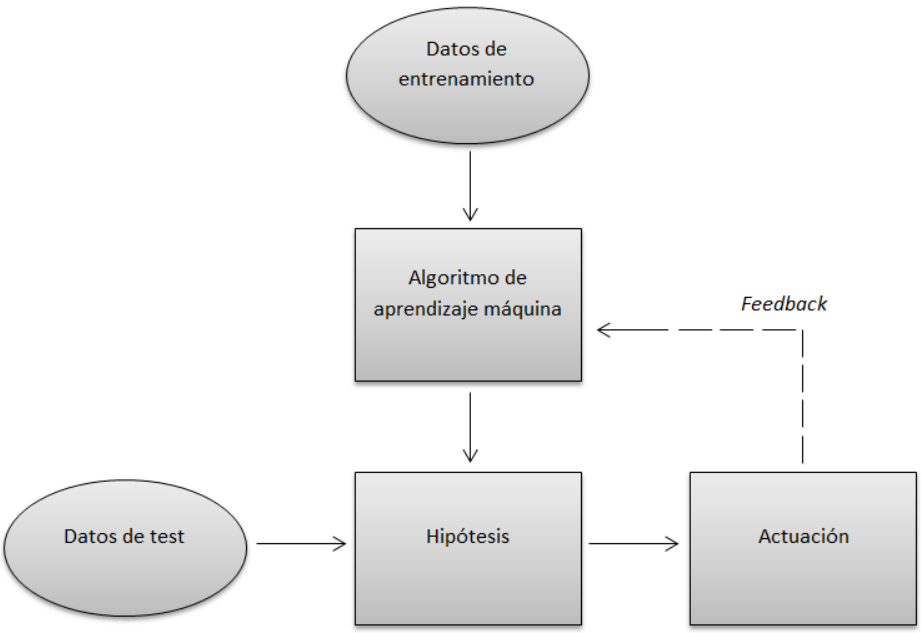


Ilustración 19. Diagrama del proceso de aprendizaje máquina

El proceso de aprendizaje máquina, mostrado en la ilustración superior, establece una hipótesis como una función entre el conjunto de datos de entrada, *x* y la salida del sistema, *y*, es decir *h*: *X→Y.* Esta función o hipótesis está construida por un conjunto de datos ya conocidos (Datos de entrenamiento) sobre los cuales el *algoritmo de aprendizaje máquina* ha sido aplicado. Una vez es construida la hipótesis/función, el siguiente paso consiste en la asignación de cada nueva entrada *x* (Datos de test) a una salida *y*, en función de la hipótesis establecida, siendo el principal objetivo de la satisfactoria clasificación que a cada nueva entrada *x* le corresponda una salida *y*.

Dependiendo de la consistencia de la hipótesis, se pueden obtener diferentes resultados, variando la actuación de nuestro sistema. De esta forma, es importante escoger un apropiado conjunto de datos de entrenamiento así como un buen algoritmo de aprendizaje que sea aplicado a dicho conjunto.

Dependiendo del caso, dos tipos de aprendizaje máquina pueden llevarse a cabo, clasificación o regresión.

La clasificación trata con el problema de predecir a qué clase de datos (de una colección ya conocida) pertenece una nueva instancia u observación, tomando como referencia un conjunto de datos de entrenamiento en el cual cada observación ya está etiquetada. El elemento a cargo de esta tarea es conocido como *clasificador*.

En cuanto al proceso de clasificación, hay varios tipos de clasificadores y algoritmos de clasificación, como clasificadores lineales, redes neuronales o árboles de decisión entre

otros. Los clasificadores que se han empleado en este proyecto son los siguientes: *ZeroR* (referencia), *Logistic*, *SimpleLogistic* y *SMO.*

4.2 WEKA

WEKA (Waikato Environment for Knowledge Analysis) [25] es una colección de algoritmos de aprendizaje máquina desarrollados en Java por la Universidad de Waikato (Nueva Zelanda) que sirven para trabajar en tareas de minería de datos. Weka contiene herramientas para pre-procesado de datos, clasificación, regresión, *clustering*, reglas de asociación, y visualización.

Este programa dispone de una licencia GPL (GNU Public License), lo que significa que es de libre distribución y difusión.

4.2.1 FORMATO DE ARCHIVO ARFF

Nativamente Weka trabaja con un formato denominado "arff" [38], acrónimo de *Attribute-Relation File Format*. Este formato está compuesto por una estructura diferenciada en tres partes [39]:

1. **Cabecera**. Se define el nombre de la relación. Su formato es el siguiente:

   @relation <nombre-de-la-relación>

   Donde *<nombre-de-la-relación>* es de tipo texto.
2. **Declaraciones de atributos**. En esta sección se declaran los atributos que compondrán nuestro archivo junto a su tipo. La sintaxis es la siguiente:

   @attribute <nombre-del-atributo> <tipo>

   Donde *<nombre-del-atributo>* es de tipo texto teniendo las mismas restricciones que el caso anterior. Weka acepta diversos tipos, pero los usados en este proyecto son dos:
   - NUMERIC. Expresa números reales.
   - ENUMERADO. El identificador de este tipo consiste en expresar entre llaves y separados por comas los posibles valores (caracteres o cadenas de caracteres) que puede tomar el atributo. En nuestro caso se introducirán los posibles valores que pueden tomar las etiquetas *valencia* y *arousal*, que serán explicadas posteriormente. Un ejemplo es el siguiente:

     @attribute 2C_VAL_LABEL {LOW_VAL, HIGH_VAL}
3. **Sección de datos**. En donde se declaran todos los datos que componen la relación separando entre comas los atributos y con saltos de línea las relaciones. Esta sección va precedida de la etiqueta *@data*.

Un ejemplo completo de archivo “arff” se muestra a continuación. Está tomado de un fichero creado para el desarrollo de este proyecto, en concreto pertenece a las características acústicas de alto nivel.

```
@relation 'Labels_TFG_Gonzalo_Solana-weka.filters.unsupervised.attribute.Remove-R2_prueba-weka

@attribute others numeric
@attribute speechmusic numeric
@attribute music numeric
@attribute speech numeric
@attribute more_than_one_events numeric
@attribute percentaje_others numeric
@attribute percentaje_speechmusic numeric
@attribute percentaje_music numeric
@attribute percentaje_speech numeric
@attribute number_of_fragments numeric
@attribute score_ot numeric
@attribute score_sm numeric
@attribute score_mu numeric
@attribute score_sp numeric
@attribute score_global numeric
@attribute only_ot numeric
@attribute only_sm numeric
@attribute only_mu numeric
@attribute only_sp numeric
@attribute 3C_ARO_LABEL {LOW_ARO,MED_ARO,HIGH_ARO}

@data
0,0,0,1,0,0,0,0,1,1,-1000,-1000,-1000,131.787005,131.787005,0,0,0,1,LOW_ARO
0,0,0,1,0,0,0,0,1,1,-1000,-1000,-1000,47.318392,47.318392,0,0,0,1,MED_ARO
1,0,1,0,1,0.647059,0,0.352941,0,2,65.334894,-1000,-31.981715,-1000,30.987856,0,0,0,0,MED_ARO
0,0,1,0,0,0,0,1,0,1,-1000,-1000,82.259271,-1000,82.259271,0,0,1,0,MED_ARO
0,0,1,0,0,0,0,1,0,1,-1000,-1000,86.340387,-1000,86.340387,0,0,1,0,MED_ARO
0,0,1,0,0,0,0,1,0,1,-1000,-1000,67.675062,-1000,67.675062,0,0,1,0,LOW_ARO
0,0,1,0,0,0,0,1,0,1,-1000,-1000,51.579382,-1000,51.579382,0,0,1,0,MED_ARO
1,0,1,0,1,0.368421,0,0.631579,0,2,0.243614,-1000,70.878784,-1000,44.8553,0,0,0,0,LOW_ARO
```

Ilustración 20. Fichero "arff"

4.2.2 WEKA EXPLORER Y WEKA EXPERIMENTER

Las herramientas que proporciona Weka usadas en este proyecto son el *Explorer* y el *Experimenter*.

Weka Explorer es una aplicación que nos permite evaluar con detalle el contenido de un fichero “arff”. Esta herramienta nos permite realizar un pre-procesado del archivo que vamos a estudiar, mostrándonos todos los atributos que contiene. Además dispone de un selector de atributos, el cual usaremos para crear subconjuntos tanto de características acústicas de bajo nivel, como de alto nivel para comprobar cuáles son las que más aportan a nuestro sistema.

Weka Experimenter, por otro lado, es una aplicación que proporciona las herramientas necesarias para probar diferentes algoritmos de clasificación sobre multitud de conjuntos de datos, con el objetivo de conseguir los resultados del proceso, que son mostrados de manera organizada.

4.3 DEFINICIÓN DE ETIQUETAS

Para poder realizar el proceso de experimentación, primero es necesario definir las etiquetas sobre las que clasificaremos nuestros datos. LIRIS-ACCEDE [1], además de los 9.800 vídeos, proporciona un ranking en donde aparecen los resultados de *valencia* y *arousal* para cada vídeo. Esta *valencia* y *arousal* son las que definen la posición de cada vídeo en el espacio 2D de *valencia-arousal* visto en la *Ilustración 8*.

Se tienen por tanto dos etiquetas sobre las que extraer información: *valencia* y *arousal*. Dividiremos cada una en 2 clases, con el objetivo de poder clasificar cada vídeo según tenga *valencia* bajo o alto y *arousal* bajo o alto. Las etiquetas son las siguientes:

- **LOW_VAL, HIGH_VAL**. Para obtener estas etiquetas se ordena de menor a mayor según el valor de la *valencia* el ranking proporcionado por LIRIS-ACCEDE y se calcula la mediana con respecto a la *valencia*. Todos los vídeos que tengan una *valencia* menor que la resultante de la mediana serán catalogados como *LOW_VAL*, y todos los vídeos que tengan una *valencia* mayor serán catalogados como *HIGH_VAL*.
- **LOW_ARO, HIGH_ARO**. Para obtener estas etiquetas se realiza el mismo procedimiento seguido anteriormente pero con la etiqueta *arousal*. Todos los vídeos que tengan un *arousal* menor que la resultante de la mediana serán catalogados como *LOW_ARO*, y todos los vídeos que tengan un *arousal* mayor serán catalogados como *HIGH_ARO*.

Para intentar ser más exactos en el ejercicio de clasificación e intentar conseguir mejores resultados, dividiremos también cada etiqueta en 3 clases, y así poder diferenciar cada vídeo entre *valencia* baja, media o alta y *arousal* bajo, medio, o alto. Resultando como sigue:

- **LOW_VAL, MED_VAL, HIGH_VAL**. Para obtener estas etiquetas se ordena de menor a mayor según el valor de la *valencia* el ranking proporcionado por LIRIS-ACCEDE. Una vez ordenados dividiremos los datos en 3 conjuntos iguales, donde cada conjunto contiene el 33% de los datos. Tras ello obtenemos el valor de *valencia* del último vídeo del primer 33% de los datos (lo llamamos *valor_1_val*) y el último valor de *valencia* del último vídeo del segundo 66% (lo llamamos *valor_2_val*). Todos los vídeos que tengan una *valencia* menor que *valor_1_val* serán catalogados como *LOW_VAL*. Todos los vídeos que tengan una *valencia* mayor que *valor_1_val* y menor que *valor_2_val* serán catalogados como *MED_VAL*. Y todos los vídeos que tengan una *valencia* mayor que *valor_2_val* se catalogarán como *HIGH_VAL*.
- **LOW_ARO, MED_ARO, HIGH_ARO**. Para obtener estas etiquetas se realiza el mismo procedimiento seguido anteriormente pero con la etiqueta *arousal* y obtenemos *valor_1_aro* y *valor_2_aro*. Todos los vídeos que tengan un *arousal* menor que *valor_1_aro* serán catalogados como *LOW_ARO*. Todos los vídeos que tengan un

arousal mayor que *valor_1_aro* y menor que *valor_2_aro* serán catalogados como *MED_ARO*. Y todos los vídeos que tengan un *arousal* mayor que *valor_2_aro* se catalogarán como *HIGH_ARO*.

Como se aprecia, hemos obtenido 4 conjuntos de etiquetas que se aplicarán a las características acústicas de alto y bajo nivel extraídas para poder realizar la experimentación.

4.4 EXPERIMENTOS REALIZADOS

Una vez que disponemos de las 392 características acústicas de bajo nivel y 19 de alto nivel para cada vídeo, además de las etiquetas resultantes de la sección anterior con las que clasificaremos cada vídeo, se realizarán los siguientes experimentos:

Características acústicas (tipo)	Etiqueta	Número de clases	Características acústicas (número)	Experimento
BAJO NIVEL	*Valencia*	2 clases: - LOW_VAL - HIGH_VAL	*Todas: 392*	**Exp. 1**
		3 clases: - LOW_VAL - MED_VAL - HIGH_VAL	*Todas: 392*	**Exp. 2**
	Arousal	2 clases: - LOW_ARO - HIGH_ARO	*Todas: 392*	**Exp. 3**
		3 clases: - LOW_ARO - MED_ARO - HIGH_ARO	*Todas: 392*	**Exp. 4**
ALTO NIVEL	*Valencia*	2 clases: - LOW_VAL - HIGH_VAL	*Todas: 19*	**Exp. 5**
		3 clases: - LOW_VAL - MED_VAL - HIGH_VAL	*Todas: 19*	**Exp. 6**
	Arousal	2 clases: - LOW_ARO - HIGH_ARO	*Todas: 19*	**Exp. 7**
		3 clases: - LOW_ARO - MED_ARO - HIGH_ARO	*Todas: 19*	**Exp. 8**

Tabla 1. Experimentos principales

Además, de cada experimento de características acústicas de bajo nivel se obtendrán las mejores 50 características y las mejores 100. Con "las mejores" nos referimos a aquellas características acústicas que más influyen en la decisión entre cada una de las clases que componen los conjuntos de etiquetas, por ejemplo decidir entre *baja* y *alta valencia* (LOW_VAL y HIGH_VAL).

También se realiza el mismo proceso con las características acústicas de alto nivel, pero en este caso seleccionando las mejores 6 y las mejores 12 características.

Hay que aclarar que las características a las que nos referimos son los atributos definidos en los ficheros "arff". Para seleccionar los mejores atributos o características se ha hecho uso de la herramienta *Weka Explorer*.

Tras esta selección de atributos/características, la suma de experimentos asciende a un total de 24. Son mostrados en las siguientes tablas:

Características acústicas (tipo)	Etiqueta	Número de clases	Características acústicas (número)	Experimento
BAJO NIVEL	*Valencia*	2 clases: - LOW_VAL - HIGH_VAL	*Todas: 392*	**Exp. 1**
			Mejores 100	Exp. 1.1
			Mejores 50	Exp. 1.2
		3 clases: - LOW_VAL - MED_VAL - HIGH_VAL	*Todas: 392*	**Exp. 2**
			Mejores 100	Exp. 2.1
			Mejores 50	Exp. 2.2
	Arousal	2 clases: - LOW_ARO - HIGH_ARO	*Todas: 392*	**Exp. 3**
			Mejores 100	Exp. 3.1
			Mejores 50	Exp. 3.2
		3 clases: - LOW_ARO - MED_ARO - HIGH_ARO	*Todas: 392*	**Exp. 4**
			Mejores 100	Exp. 4.1
			Mejores 50	Exp. 4.2

Tabla 2. Experimentación completa para "características acústicas de bajo nivel"

Características acústicas (tipo)	Etiqueta	Número de clases	Características acústicas (número)	Experimento
ALTO NIVEL	*Valencia*	2 clases: - LOW_VAL - HIGH_VAL	*Todas: 19*	**Exp. 5**
			Mejores 12	Exp. 5.1
			Mejores 6	Exp. 5.2
		3 clases: - LOW_VAL - MED_VAL - HIGH_VAL	*Todas: 19*	**Exp. 6**
			Mejores 12	Exp. 6.1
			Mejores 6	Exp. 6.2
	Arousal	2 clases: - LOW_ARO - HIGH_ARO	*Todas: 19*	**Exp. 7**
			Mejores 12	Exp. 7.1
			Mejores 6	Exp. 7.2
		3 clases: - LOW_ARO - MED_ARO - HIGH_ARO	*Todas: 19*	**Exp. 8**
			Mejores 12	Exp. 8.1
			Mejores 6	Exp. 8.2

Tabla 3. Experimentación completa para "características acústicas de alto nivel"

Los resultados de estos 24 experimentos son mostrados y estudiados en el siguiente capítulo.

CAPÍTULO 5. RESULTADOS

El objetivo de este capítulo es mostrar y analizar los resultados obtenidos por WEKA tras los 24 experimentos realizados. Dichos resultados se estudiarán en dos partes de forma independiente: primero para las características acústicas de bajo nivel y después para las características acústicas de alto nivel. Después se hará una selección con los mejores resultados obtenidos y se combinarán características de alto y bajo nivel. Por último, se estudiarán los resultados obtenidos de la experimentación conjunta de características.

Para agrupar de forma clara los experimentos realizados en las tablas que se mostrarán en este capítulo se ha creado una notación, formada por 3 componentes:

- **Número de clases**: podrá ser *2C* o *3C*. En el caso de 2C se corresponderá a una etiqueta (*valencia* o *arousal*) bajo o alto. En el caso de 3C se corresponderá a una etiqueta (*valencia* o *arousal*) bajo, medio o alto.

- **Etiqueta**: podrá ser *val* (*valencia*) o *aro* (*arousal*).

- **Número de características empleadas**: podrá ser *all* (todas), y *top100* (mejores 100) o *top50* (mejores 50) para características acústicas de bajo nivel, y *top12* (mejores 12) o *top6* (mejores 6) para características de alto nivel.

Por ejemplo, para el caso en el que estemos frente a un experimento de *valencia*, con 2 clases, y se empleen las mejores 50 características la notación sería la siguiente: *2C_val_top50.*

Además, se han pintado de color naranja los experimentos correspondientes a la *valencia* y de azul los correspondientes al *arousal*. El asterisco indica el valor más alto para el experimento general llevado a cabo.

5.1 CARACTERÍSTICAS ACÚSTICAS DE BAJO NIVEL

A continuación se presentan los resultados obtenidos de la experimentación para cada clasificador con características acústicas de bajo nivel.

Experimento	ZeroR	Logistic	SimpleLogistic	SMO
Exp. 1 (2C_val_all)	54.80	54.07	55.54	55.63
Exp. 1.1 (2C_val_top100)	54.80	56.63	56.32	56.26
Exp. 1.2 (2C_val_top50)	54.80	56.50	57.82 *	56.40
Exp. 2 (3C_val_all)	39.50	36.79	40.22	40.56
Exp. 2.1 (3C_val_ top100)	39.46	51.35 *	49.06	50.37
Exp. 2.2 (3C_val_ top50)	39.46	49.69	49.27	48.30
Exp. 3 (2C_aro_all)	52.80	55.08	59.32	59.87
Exp. 3.1 (2C_aro_ top100)	52.80	61.19	61.54	62.77 *
Exp. 3.2 (2C_aro_ top50)	52.80	61.03	61.86	62.29
Exp. 4 (3C_aro_all)	34.80	35.53	40.03	40.24
Exp. 4.1 (3C_aro_ top100)	34.80	50.16	49.14	50.27 *
Exp. 4.2 (3C_aro_ top50)	34.80	48.90	48.64	49.25

Tabla 4. Resultados para "características acústicas de bajo nivel"

Para la *valencia*, en el caso de 2 clases se obtiene un 57.82% frente a un 54.80% de la referencia (*ZeroR*). Para el caso de 3 clases se obtiene un 51.35% frente a un 39.46%. Como podemos observar, la *valencia* mejora aproximadamente un 11% al usar 3 clases (*valencia* alta, media y baja), mientras que al usar 2 clases solo se mejora un 3%.

Sin embargo, para el *arousal* el sistema mejora sustancialmente empleando 2 o 3 clases. En el caso de 2 clases se obtiene un 10% de mejora frente a la referencia, mientras que el caso de 3 clases se consigue un 15% aproximadamente.

Un detalle que se debe apreciar es que se obtienen mejores resultados de los clasificadores usando un subconjunto de características acústicas en lugar de usar todas las disponibles (392). Como vemos en la tabla superior, 3 de los 4 asteriscos (valores más altos) se dan en experimentos realizados con la selección de las 100 mejores características acústicas de bajo nivel. Esto es debido al problema de sobreajuste que ofrece el aprendizaje máquina.

5.2 CARACTERÍSTICAS ACÚSTICAS DE ALTO NIVEL

En la siguiente tabla se muestran los resultados obtenidos de la experimentación para cada clasificador de las características acústicas de alto nivel.

Experimento	ZeroR	Logistic	SimpleLogistic	SMO
Exp. 5 (2C_val_all)	50.00	52.36	52.59	51.79
Exp. 5.1 (2C_val_top12)	50.00	52.75	52.57	52.44
Exp. 5.2 (2C_val_top6)	50.00	52.92 *	52.52	52.66
Exp. 6 (3C_val_all)	33.35	35.82	35.76	35.46
Exp. 6.1 (3C_val_top12)	33.35	35.83	35.76	35.46
Exp. 6.2 (3C_val_top6)	33.35	36.35 *	35.88	35.94
Exp. 7 (2C_aro_all)	50.00	52.77 *	52.24	51.53
Exp. 7.1 (2C_aro_top12)	50.00	52.59	51.77	51.82
Exp. 7.2 (2C_aro_top6)	50.00	52.09	51.72	51.29
Exp. 8 (3C_aro_all)	33.35	36.18	35.87	34.82
Exp. 8.1 (3C_aro_top12)	33.35	36.14	35.84	35.48
Exp. 8.2 (3C_aro_top6)	33.35	36.67 *	36.07	35.89

Tabla 5. Resultados para "características acústicas de alto nivel"

Se puede ver, tanto para la *valencia* como para el *arousal*, y usando 2 o 3 clases, que el sistema mejora un 3% aproximadamente.

Esta cifra, aunque es baja, es significativa. Hay que recordar que se han propuesto 19 características acústicas de alto nivel, y que son enfrentadas una base de datos muy grande. Es por ello, que proporcionalmente, el 3% de mejora conseguido nos ayuda a entender algo mejor cómo pueden afectar estas características propuestas en un espectador al visualizar un vídeo.

5.3 ANÁLISIS CONJUNTO

En esta sección se experimentará y se analizará de forma conjunta las características acústicas de bajo nivel con las de alto nivel. Para ello se trabajará solo con 2 clases, es decir, baja o alta *valencia* y bajo o alto *arousal*.

Se realizan dos nuevos experimentos: Experimento 9 y Experimento 10.

Para la *valencia*, se combina el archivo con las mejores 50 características acústicas de bajo nivel con el archivo que contiene las mejores 6 características acústicas de alto nivel. Se obtiene por tanto un nuevo fichero de características que denominaremos *val_top6_top50*. Este experimento será nombrado como *Exp. 9*.

Para el *arousal* se realiza el mismo proceso. El nuevo fichero de características se llamará *aro_top6_top50*. Este experimento se nombrará como *Exp. 10*.

Los resultados se muestran a continuación:

Experimento	ZeroR	Logistic	SimpleLogistic	SMO
Exp. 9 (val_top6_top50)	54.80	56.60	56.93	57.15 *
Exp. 10 (aro_top6_top50)	52.80	60.31	61.45	62.89 *

Tabla 6. Resultados de la combinación de características acústicas de alto y bajo nivel

Si consultamos los resultados conseguidos anteriormente en la *sección 5.1* y en la *sección 5.2* para 2 clases, la *valencia* mejoraba un 3% y el *arousal* mejoraba el sistema en un 15% únicamente en el caso de las características acústicas de bajo nivel.

Viendo los resultados obtenidos en los experimentos 9 y 10, podemos ver que para la *valencia* existe una mejora algo menor que 3% con respecto a la referencia y que para el *arousal* la mejora es del 10%. Por tanto, realizar experimentos combinando características acústicas de bajo y alto nivel no mejora los resultados que ya habíamos conseguido de forma independiente.

Como última comprobación, se hará una selección de los mejores atributos que conforman los experimentos 9 y 10. Para ello se seleccionarán las mejores 40, mejores 30, mejores 20 y mejores 10 características acústicas de dichos experimentos.

Primero se muestra una tabla de los experimentos que surgen del *Exp. 9*, el cual corresponde a la combinación de características acústicas de alto y bajo nivel para la etiqueta de *valencia*. Tras esto, se analizan los resultados y se mostrará otra tabla con los

experimentos surgidos de la selección de atributos del *Exp. 10*, el cual corresponde a la combinación de características acústicas de alto y bajo nivel para la etiqueta de *arousal*.

Experimento	ZeroR	Logistic	SimpleLogistic	SMO
Top 10 de Exp. 9	54.80	59.79	59.35	60.25 *
Top 20 de Exp. 9	54.80	58.90	58.45	59.60
Top 30 de Exp. 9	54.80	58.98	58.10	59.73
Top 40 de Exp. 9	54.80	58.28	58.00	59.33

Tabla 7. Resultados de la selección de las mejores características del Exp. 9

El objetivo es mejorar el 57.15% obtenido en el *Exp. 9* para la *valencia* usando todas las características resultantes de la combinación de las de alto y bajo nivel. Vemos que cualquier selección de las mejores características acústicas mejora el resultado anterior. En el mejor de los casos, la selección de las mejores 10 características mejora el resultado hasta en un 3% con respecto a usar todas las características disponibles, siendo un 6% con respecto a la referencia (54.80%).

Experimento	ZeroR	Logistic	SimpleLogistic	SMO
Top 10 de Exp. 10	52.80	62.92	62.48	62.95
Top 20 de Exp. 10	52.80	64.40	63.72	64.14
Top 30 de Exp. 10	52.80	63.97	63.54	64.01
Top 40 de Exp. 10	52.80	62.95	62.61	64.43 *

Tabla 8. Resultados de la selección de las mejores características del Exp. 10

Para el arousal, el objetivo es mejorar el 62.89% obtenido en el *Exp. 10*. Se mejora en aproximadamente un 2% al usar las mejores 40 características resultantes de la combinación de las de alto y bajo nivel, siendo en total un 12% con respecto a la referencia (52.80%).

CAPÍTULO 6. GESTIÓN DEL PROYECTO

En este capítulo se integra la organización que se ha llevado a cabo para desarrollar el proyecto, así como el número de horas empleadas y un presupuesto estimado para el mismo.

6.1 ORGANIZACIÓN

Las etapas en las que se ha dividido el proyecto son las mostradas en el diagrama de la siguiente página (ver *Ilustración 21*). Hay que indicar que la redacción de la memoria del proyecto no ha sido incluida en el diagrama de organización.

El tiempo empleado en cada una de las etapas se desglosa en la siguiente tabla:

Etapas	Horas empleadas
Investigación previa	20
Planteamiento del problema	20
Estudio de las herramientas	30
Extracción y proposición de características acústicas	70
Pruebas, correcciones y mejoras	60
Experimentación	40
Análisis de los resultados y conclusiones	30
Redacción de la memoria	70
Total	**340**

Tabla 9. Desglose de horas

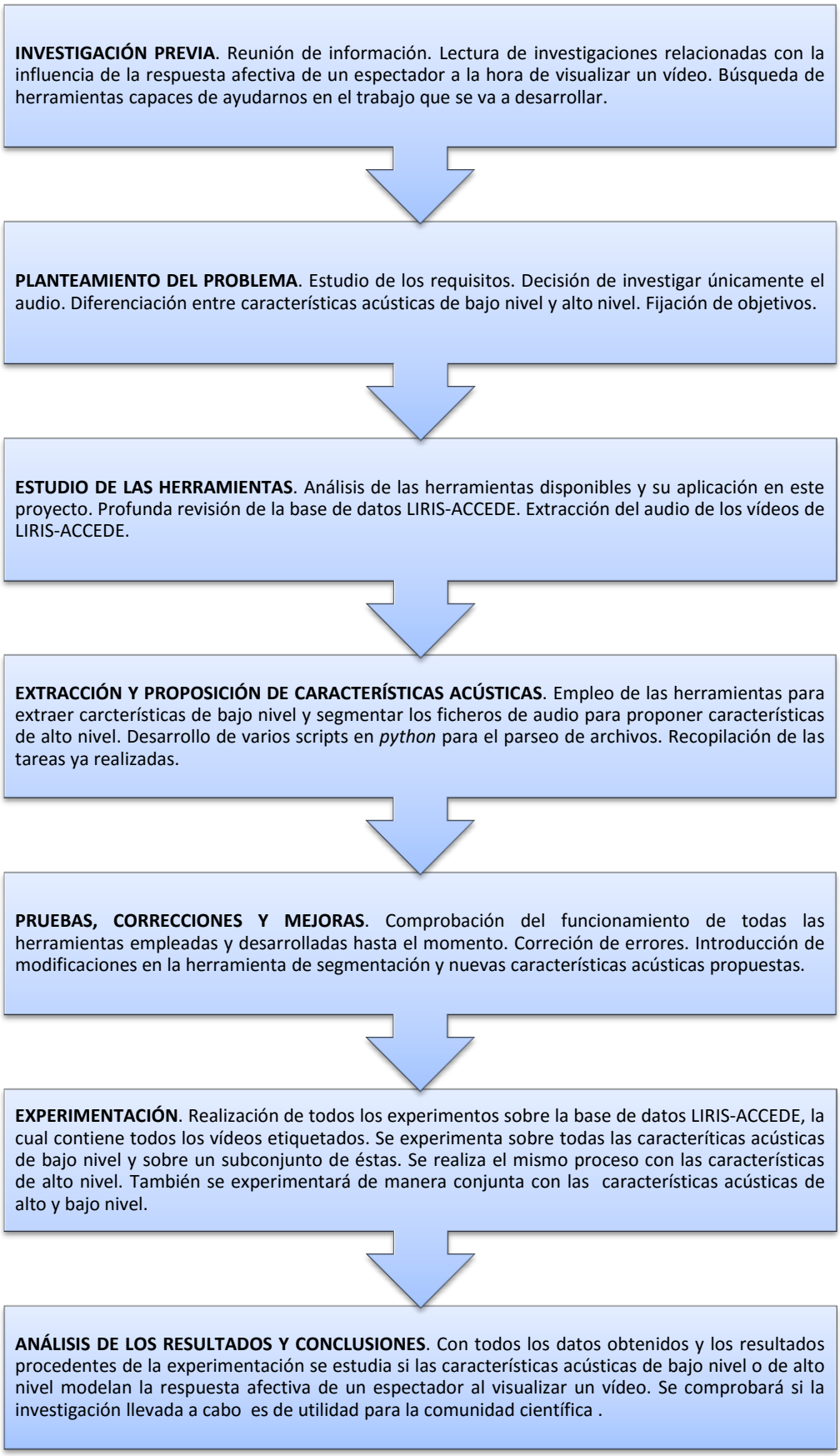


Ilustración 21. Diagrama de organización del proyecto

6.2 PRESUPUESTO

El objetivo de este apartado es estimar el coste total del proyecto, tanto del material empleado como de personal.

El coste material es el conformado por el equipo que ha sido usado para el desarrollo del proyecto y por las licencias de software requeridas. Se realiza un desglose del precio, incluyendo el periodo de amortización del mismo, y calculando el coste en los aproximadamente 6 meses que tiene de duración el proyecto.

Material	Precio	Amortización	Coste
Laptop Asus N53S	550€	5 años	30€
Ampliación 4Gb RAM	33€	5 años	1,8€
SSD 128Gb	130€	6 años	5€
Windows 7	-	-	-
MATLAB R2015b	69€	1 año	34,5€
Microsoft Office 2013 Home	119€	5 años	11,9€
SoX	-	-	-
MIRtoolbox	-	-	-
HTK	-	-	-
openSMILE	-	-	-
WEKA	-	-	-
Python(x,y)	-	-	-
Total	**901€**	**-**	**83,2€**

Tabla 10. Coste del material

Hay que mencionar que la versión de Windows 7 es ofrecida de manera gratuita por la Universidad Carlos III de Madrid. El resto de material que no dispone de precio es el software de libre distribución empleado.

A continuación se muestra el coste del personal:

Puesto	Precio/hora	Horas	Coste total
Jefe de proyecto	25€/hora	50	1.250€
Ingeniero	7€/hora	340	2.380€
Total	**-**	**-**	**3.630€**

Tabla 11. Coste de personal

Por último, calculamos el precio total del proyecto:

Material	83,2€
Personal	3.630€
Total	**3.713,2€**

Tabla 12. Coste total del proyecto

CAPÍTULO 7. CONCLUSIONS AND FUTURE WORK

This chapter summarizes the most important aspects of this study, as well as some lines of work that can be implemented in the future to improve the results.

7.1 CONCLUSIONS

The objective of this study is to analyze the influence of audio in the affective response of a spectator while he or she visualizes a video. High and low level acoustic characteristics have been extracted and afterwards proposed to help understand, in an objective way, how audio is capable of affecting people's emotions.

The starting point of the project is possible due to the free access to the LIRIS-ACCEDE database [1] offered to the scientific community. The database provides a ground truth: 9,800 video extracts, labeled in accordance with *valence* and *arousal* and with a great diversity of contents, genres and languages.

Once the data is obtained from LIRIS-ACCEDE, the audio is separated from each video of the database so it is possible to obtain the characteristics.

MIRToolbox [2] is the tool that has made the extraction of the low level characteristics from the audio files possible. Each audio file has yielded 392 statistical characteristics.

In order to extract high level acoustic characteristics, a previous phase in which the audio files were segmented into different acoustic categories was necessary: *speech*, *speech-music*, *music* and *others*. The segmentation was executed using a tool supplied by A. Gallardo and R. San Segundo [8]. This tool is made up of models trained to segment audios belonging to a database of news in Catalan. The use of this tool to process our LIRIS-ACCEDE data has generated a high average error rate compared to the results of other segmentations.

The rate is high for the following two reasons: (i) the audio files are very diverse but they have been segmented with a tool whose models have been trained using only files in Catalan, and (ii) the segmentation has been evaluated using a reference file labeled by one person and which contains only 100 items (in contrast to the 9.800 files of the LIRIS-ACCEDE database). However, the rate has been considered acceptable and thus adequate to obtain high level acoustic characteristics, making it possible to continue with the study.

After the analysis of the segmentation tool and the results, 19 high level acoustic characteristics have been proposed. To obtain the characteristics, a Python script was created. The script is capable of extracting data that belong to each audio file resulting from the segmentation.

Finally, a test was carried out to ascertain whether the high and low level characteristics influence the affective response of the spectator. The test consists of an experimentation process using valence and arousal labels. During the classification process, 2 or 3 classes for each label were used:

- When 2 classes are used, both *valence* and *arousal* may have a *low* or *high* value.
- When 3 classes are used, both *valence* and *arousal* may have a *low*, *medium* or *high* value.

The results of the experimentation, shown in *Chapter 5*, demonstrate that the proposed high level acoustic characteristics slightly improve the system by 3% for both *arousal* and *valence*, and using 2 or 3 classes. Nevertheless, low level characteristics improve the system by 15% for *arousal* and by 11% for *valence*. In the latter, 3 classes were used, while using 2 generates good results for *arousal* but unsatisfactory ones for *valence*.

Subsequently, further experimentation was carried out. This time, high and low level acoustic characteristics were jointly tested, using *valence* and *arousal* labels with only 2 classes: *low* and *high*. The outcome of this experimentation does not improve the results of experimentation employing only low level acoustic characteristics.

Therefore, a selection has been made of the best characteristics that arise from the combination of high and low level acoustic characteristics. The outcome of this selection has proved satisfying, obtaining a 14% improvement with respect to the reference. The improvement is due to the overfitting that takes place in machine learning.

Ultimately, the study has yielded three main contributions:

- First would be that the proposed high level acoustic characteristics somewhat improve the system (by 3%) in comparison with the reference. Although it is a low percentage, it is significant. The reason being that 19 high level acoustic characteristics is a very low number compared to the great amount of data in LIRIS-ACCEDE, and thus, too few to solve a problem of the scale of this study. Therefore, even though the improvement is small, it is an important step towards understanding the influence of audio in the field of study of affective response.

- The second contribution is related to the obtained low level acoustic characteristics. These characteristics, which in total are 392, are indeed significant in comparison to the data from LIRIS-ACCEDE. Furthermore, after experimenting with a selection of

the 100 best characteristics, the results are fruitful: the best score achieved is 15% higher than the one offered by the reference system. This means that low level acoustic characteristics are of a great importance, and they have a relevant influence on the affective response of the spectator

- Finally, it is worth considering the influence that the high and low level acoustic characteristics have on the affective response of the spectator when acting jointly. The last part of the project has proved the importance of studying these characteristics together, as the results have been substantial (14% of improvement with respect to the reference). The consequent line of work could be relevant in many domains and may be of interest for the scientific community that studies the affective response of the individuals during a multimedia event.

7.2 FUTURE WORK

Below are described the future lines of work, as well as improvements that can be implemented in this project in order to obtain superior results and to perform a more detailed analysis of the affective response:

- Joint study of *valence* and *arousal*. A possible enhancement of the results could follow from the joint experimentation in the *valence-arousal* space. This exercise could include combined labels for each video such as *low arousal* and *low valence*, *low arousal* and *high valence*, *high arousal* and *low valence*, *high arousal* and *high valence*.

- Study of the images that form the video. Throughout this research, only the audio signal of the video has been considered. It may be of interest to introduce visual characteristics based on the images from the video to analyze visual and audio characteristics together and their influence on the affective response.

- Develop a segmentation tool that can deal with a very large database such as LIRIS-ACCEDE. In this way, the average error rate from the segmentation evaluation would improve and, therefore, the obtained high level acoustic characteristics would be of better quality.

- Provide new tools for the obtaining of more specialized high level acoustic characteristics. For example, a system capable of differentiating how many interlocutors appear in a voice fragment or of identifying the sex of the person who is talking.

- Improve the reference file used during the segmentation evaluation. For this matter, it is convenient to (i) label a larger set of videos than the one used in this project and (ii) have more than one person labeling each video. In this way, segmentation evaluation could be deeper, the reference file being a great set of reliable data with which to compare the file obtained from the segmentation.

- Research of the affective response based on the continuous analysis of the affective content of a video. This system estimates an affective score for each frame of the video. To perform this study based on continuous analysis, a previous study of the images that structure the video is required.

APÉNDICES

Esquema de las características estadísticas disponibles en la herramienta MIRtoolbox:

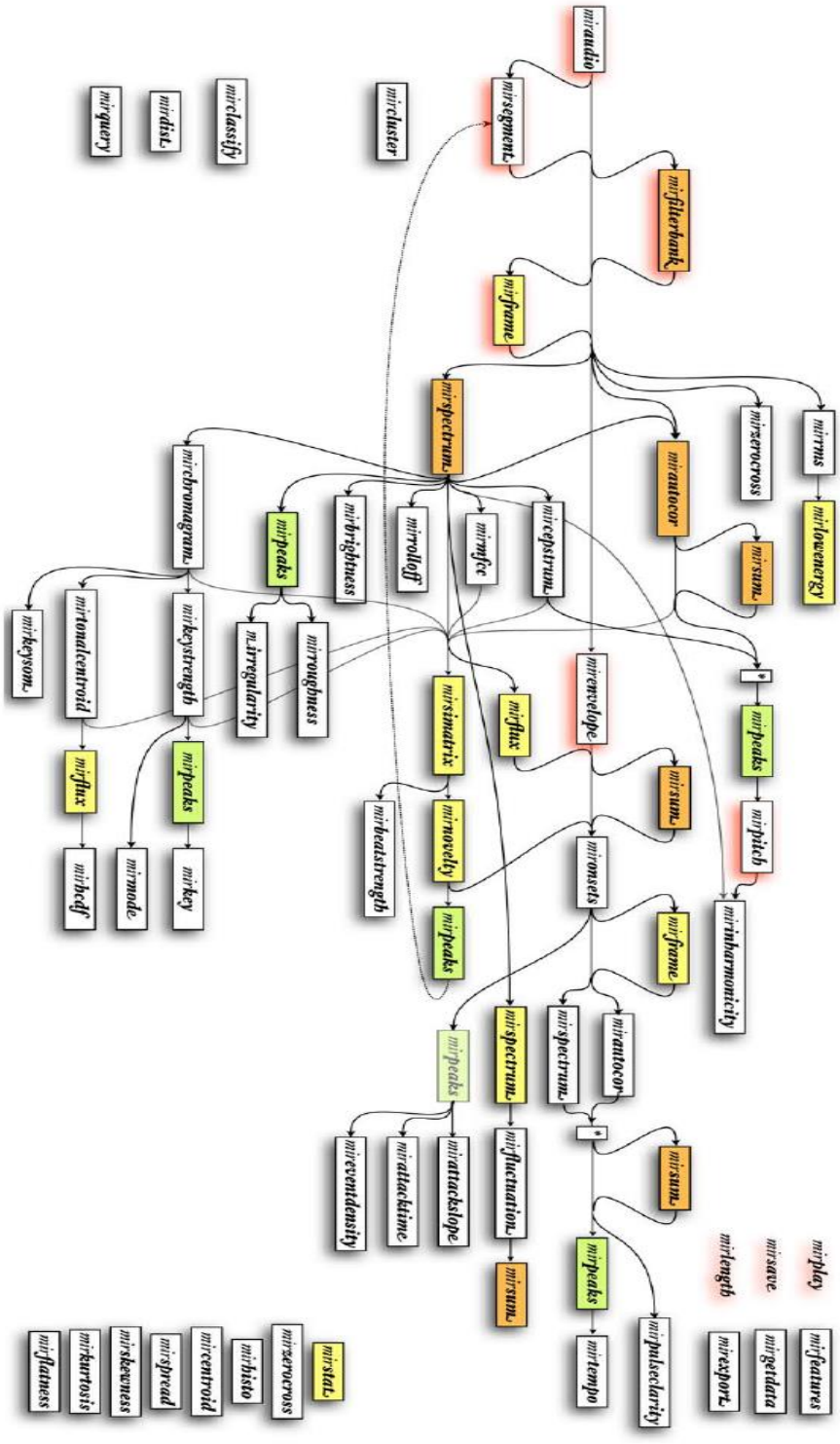


Ilustración 22. Características disponibles en MIRtoolbox

REFERENCIAS Y BIBLIOGRAFÍA

[1] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, “LIRIS-ACCEDE: A Video Database for Affective Content Analysis,” in *IEEE Transactions on Affective Computing*, 2015.

[2] Olivier Lartillot and Petri Toiviainen, “A MATLAB toolbox for musical feature extraction from audio”, *10th Int. Conference on Digital Audio Effects (DAFx-07),* September 2007.

[3] A. Hanjalic and L.-Q. Xu, “Affective video content representation and modeling,” *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 143–154, Feb. 2005.

[4] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, “A supervised approach to movie emotion tracking,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 2376–2379.

[5] H.-B. Kang, “Affective content detection using HMMs,” in *Proceedings of the eleventh ACM international conference on Multimedia*, ser. MULTIMEDIA ’03, 2003, pp. 259–262.

[6] H. L. Wang and L.-F. Cheong, “Affective understanding in film,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 6, pp. 689–704, Jun. 2006.

[7] K. Sun and J. Yu, “Video affective content representation and recognition using video affective tree and hidden markov models,” in *Affective Computing and Intelligent Interaction*, 2007, vol. 4738, pp. 594–605.

[8] A. Gallardo and R. San Segundo, “UPM-UC3M system for music and speech segmentation”, *Albayzín evaluation 2010 on Audio Segmentation,* 2010.

[9] T. Butko, C. Nadeu and H. Shulz, “Albayzin-2010 Audio Segmentation Evaluation: Evaluation Setup and Results”, *FALA 2010 - VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop,* 2010.

[10] Schulz, Henrik / Fonollosa, José A. R. (2009): "A Catalan broadcast conversational speech database", In *SLTECH-2009*, 27-29. http://www.isca-speech.org/archive_open/sltech_2009/isl9_027.html

[11] Tomonori Izumitani, Ryo Mukai, and Kunio Kashino.” A background music detection method based on robust feature extraction”. Proc ICASSP 2008.

[12] Ascensión Gallardo-Antolín and Juan M. Montero, “Histogram Equalization-Based Features for Speech, Music, and Song Discrimination”. *IEEE Signal processing letters*, Vol. 17, No. 7, 2010.

[13] Jitendra Ajmera, Iain McCowan, Herve Bourlard. “Speech/music segmentation using entropy and dynamism features in a HMM classification framework”. *Speech Communication 40* (2003) 351–363.

[14] Costas Panagiotakis and George Tziritas, “A Speech/Music Discriminator Based on RMS and ZeroCrossings”. *IEEE Trans. On Multimedia*, Vol. 7, No. 1, Feb 2005.

[15] Yizhar Lavner1 and Dima Ruinskiy, “A Decision-TreeBased Algorithm for Speech/Music Classification and Segmentation”. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009.

[16] Alessandro Bugatti, Alessandra Flammini, PierangeloMigliorati “Audio Classification in Speech and Music: A Comparison Between a Statistical and a Neural Approach”. EURASIP Journal on Applied Signal Processing 2002:4, 372–378.

[17] Mateu Aguilo, Taras Butko, Andrey Temko, Climent Nadeu “A Hierarchical Architecture for Audio Segmentation in a Broadcast News Task”. Proc I SLtech 2009. Lisbon.

[18] Cemil Demir; Erdem Ünal, Mehmet Ugur Dogan, “A Sphinx Based Speech-Music Segmentation Front-End For Improving The Performance Of An Automatic Speech Recognition System In Turkish”. CMU Sphinx Workshop 2010.

[19] How Content ID works https://support.google.com/youtube/answer/2797370?hl=en

[20] How does Siri work? http://www.quora.com/How-does-Siri-work

[21] System Requirements for MATLAB & Simulink R2015b http://es.mathworks.com/support/sysreq/current_release/

[22] Florian Eyben, Felix Weninger, Florian Gross, Björn Schuller: "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor", *In Proc. ACM Multimedia (MM)*, Barcelona, Spain, ACM, ISBN 978-1-4503-2404-5, pp. 835-838, October 2013. http://www.audeering.com/research/opensmile

[23] S. Young et al., HTK-Hidden Markov Model Toolkit. Cambridge, MA: Cambridge Univ., 2002. http://htk.eng.cam.ac.uk/

[24] Python(x,y) http://python-xy.github.io/

[25] WEKA http://www.cs.waikato.ac.nz/ml/weka/

[26] SoX – Sound eXchange http://sox.sourceforge.net/

[27] M.M. Bradley and P.J. Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994.

[28] K.R. Scherer. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729, 2005.

[29] T. Banziger, V. Tran, and KR Scherer. The geneva emotion wheel: A tool for the verbal report of emotional reactions. *International Society for Research on Emotion*, 2005.

[30] R. Cowie, G. McKeown, and E. Douglas-Cowie. Tracing emotion: an overview. *International Journal of Synthetic Emotions (IJSE)*, 3(1):1–17, 2012.

[31] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder. 'feeltrace': An instrument for recording perceived emotion in real time. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[32] N.H. Frijda. *The emotions*. Cambridge University Press, 1987.

[33] Creative Commons http://creativecommons.org/

[34] CrowdFlower http://www.crowdflower.com/

[35] R. B. Dietz and A. Lang, "Aefective agents: Effects of agent affect on arousal, attention, liking and learning," in *Cognitive Technology Conference*, 1999.

[36] J.A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[37] Youtube Statistics http://www.youtube.com/yt/press/es/statistics.html

[38] New Zealand University of Waikato. Attribute-relation file format (arff). http://www.cs.waikato.ac.nz/~ml/weka/arff.html, Julio 2013.

[39] Diego García Morate. Manual de WEKA. diego.garcia.morate(at)gmail.com

EXTENDED ABSTRACT

Internet is commonly used as a digital platform containing information has been the booster of a technological revolution. Internet grants access to an extensive number of data bases containing large amounts of information, including images, videos or audio files. Millions of users access daily many different platforms to watch, download or upload these videos.

Videos are composed by audio and a combination of images, which appear as a succession of one after another every few millisecond. Furthermore, by only studying the characteristics of the audio, contained in the video, it is possible to obtain a diverse and useful amount of information.

Therefore, an interesting investigation to this field of study is to research the influence audio has on the affective response of a spectator when viewing a video. In order to correctly study this, low level acoustic characteristics will be withdrawn and a new set of high level acoustic characteristics will be proposed.

Index Terms: audio information retrieval, affective response, emotion recognition, audio feature extraction.

STATE OF THE ART

The researches regarding affective video analysis can be classified into two subgroups:

- A **continuous** analysis of the affective content in the video, which predicts an affective score for each frame of the video. Such investigations have as an aim the mapping of the characteristics of a video in a bi-dimensional space: *valence* and *arousal*.
- A **discrete** analysis of the affective content in the video, which predicts an affective score for each segment of the video and its main aim is to achieve a category for each section of the video. Some of these categories could be “happiness”, “anger” and “sadness”.

Since the aim of this project is to extract acoustic characteristics, it is important to observe the segmentation of audio files. So in order to develop a tool with such aim, many previous studies have been based on the **analysis of statistic characteristics** from the audio signal, in order to distinguish between voice and music. On the other hand, another technique used is

based on the **architecture of the system**, in which decision algorithms based on a tree structure are proposed to achieve a segmentation between different acoustic events.

The segmentation tool that will be used in this project is the one created by A.Gallardo and R. San Segundo. Such tool was created as a design to the audio segmentation competition Albayzin 2010, in which it obtained the highest rank of the evaluation. The method used is based on *Hidden Markov Models* (HMMs), including one HMM of three stages for each acoustic category ("voice", "music", "voice over noise" and "voice over music").

The growth of the study field based on the analysis of the affective content of a video is highly significant as it helps with the creation of a large number of new or improved applications, which vary in a large range going from security systems based on voice recognition to delivery of personalized contents to the users. In a more specific field, an implementation which is under a continuing development and improvement are the copyright control systems. The use of comparing multimedia content is highly useful if it is efficient as it automatizes a costly process, which if made by an individual would take a much longer time.

IMPLEMENTATION OF A TECHNICAL SOLUTION

The research in this project has as an aim to observe the influence of audio in the affective response of a viewer when watching a video. In order to do so, the tool MIRtoolbox will allow to study the lower level acoustic characteristics along with the ones that have a higher influence on the affective response of the viewer. On the other hand, higher level acoustic characteristics obtained from the segmentation of audio files will be recommended and its influence on the viewers will be studied.

From doing so, the results obtained will allow us to observe whether the existing lower level acoustic characteristics along with the new recommended higher level acoustic characteristics model significantly the affective response of the viewer, or whether we could reject this type of characteristics as they are not relevant.

Previous research

The scale used throughout this project is *valence-arousal*, which is based on a continuous model of emotions and is commonly used in researches related to the topic of affection. This scale is based on the idea that each emotional stage can be plotted in a two-dimensional plane, being *valence* and *arousal* the two main axes. The *arousal* axis could vary from inactive (bored, passive) to active (enthusiastic). On the other hand the *valence* axis ranges from unpleasant (sad, stressed) to pleasant (happy, ecstatic).

The video database used for the development of this research is the LIRIS-ACCEDE database, and its main characteristics are the following:

- It is composed by 9800 excerpts (between 8 and 12 seconds) obtained from 160 different films (classified according to 9 genders) and short-films. Furthermore, its main language is English; however there is as well a small group of other languages such as French, German, Hindu, Spanish or Italian.
- It is available to the scientific community and shared under the license of Creative Commons.
- It is the largest data based composed by videos which represents the majority of the population, and in which emotional tags are used. All of the excerpts are categorized through crowdsourcing (open collaboration) in the two-dimensional space of *valence-arousal*.

The characteristics mentioned above allow for the database to be relevant for our study. The database is very large and its contents very diverse, making available a ground truth with reliable data with which to work.

In order to obtain the different characteristics of higher and lower level, it is necessary to initially extract the audio from each video in the LIRIS-ACCEDE database. In order to extract the audio, the tool MIRtoolbox is used all resulting audio files are stored in a directory for its future use.

Low-level acoustic characteristics

Low level acoustic characteristics are those that correspond to the result of a basic processing of the audio signal through MIRtoolbox (spectrum, autocorrelation, standard deviation, energy of the signal, zero crossings, etc.).

The use of the audio processor MIRtoolbox allows to calculate 392 acoustic characteristics for each previously obtained audio file. The above mentioned set of acoustic characteristics are labeled as lower level acoustic characteristics.

The result of the characteristics extraction process is 9800 files with the extension “arff” which will be stored in a directory for its following use.

High-level acoustic characteristics

High level acoustic characteristics are defined as acoustic characteristics which can be suggested once the audio file has surpassed a segmentation stage.

The main aim of a segmentation stage is to perceive the different acoustic events which are comprised in an audio file, and moreover find the duration of each event within the file. The acoustic events which will be acknowledged are *speech*, *music*, *speech-music* and *others*.

The tool used throughout this research is the one provided by A. Gallardo and R. San Segundo, which was created for the audio segmentation competition Albayzin 2010. This

tool has been designed in order to divide into segments audios from a news channel, which were obtained from a different database than LIRIS-ACCEDE. The audios used were 83% in Catalan and the rest in Spanish and therefore the models (HMMs) have been created under these conditions. After using this tool for the audio files used in this research (obtained from the data base LIRIS-ACCEDE), the next step is to evaluate the quality of the segmentation which has been done.

The mentioned evaluation yields a very high error ratio (approx. 80%), which indicates that the segmentation carried out is not significant enough to obtain high level acoustic characteristics from it.

As a result of it, several adjustments are made to the segmentation tool, which leads to a decrease in the error rate to a 50%. The new value obtained is still relatively high, however it is sufficient to continue with the research and propose high level characteristics from the segmentation done.

The fact that the error is still high but reasonable enough to continue with the research is due to the diversity of acoustic events contained in the LIRIS-ACCEDE data base, which does not coincide with the data base from which the segmentation tool was created. The data base used to create the segmentation tool is composed by audios which are not diverse and coming from mainly news channels. While the audios from the LIRIS-ACCEDE are obtained from films in different languages, in which the director of each film decide which is the reaction that they expect to create in the spectator with different sound effects, which complicates the segmentation.

Once the error rate is enough to continue, 19 high level characteristics are proposed which will be used to verify if they have an effect on the affective response of a viewer when watching a video or whether they do not contribute information which can be used for further research.

WEKA EXPERIMENT AND RESULTS

WEKA is a powerful collection of machine learning algorithms for data mining tasks. The use of WEKA is noteworthy in order to evaluate the statistic relevance of our data, and its use in the classification and prediction tasks.

The research has been done in order to verify whether, the high level acoustic characteristics along with the low level acoustic characteristics, have a significant influence on the affective response of a viewer. In order to do so, the tags *valence* and *arousal* has been used. Furthermore, for the classification process, 2 and 3 classes for each tag have been adopted. This means that for both tags, *valence* and *arousal*, if 2 classes have been used, the values that can be obtained are either high or low, while if 3 classes have been used the possible values attained can be low, medium or high.

The results obtained regarding the low level acoustic characteristics are favorable, as they exceed in approximately 15% the reference value of the *ZeroR* classifier. On the other hand, the results obtained regarding the high level acoustic characteristics are not as satisfying, since the improvements obtained are of 3% as a comparison to the reference value.

Additionally, a research has been conducted for the high and low acoustic characteristics combined together. The results obtained, have been positive as the system exceeds the reference level by 14%.

To conclude, the 3 experiments carried out yield positive results, which confirm that both the high level acoustic characteristics and the low level ones, obtained from the videos of the LIRIS-ACCEDE database, are significant to the understanding of the modeling of an affective response from a viewer when watching a video. Therefore, the initial aim of this project is fulfilled.

SCHEDULE AND BUDGET

The research for this project has been done from a personal computer, which limits the results obtained. This is because of the high computational cost that implies working and experimenting with all of the 9800 videos included in the data base LIRIS-ACCEDE. Therefore, if the same study had been done with a processor which had a higher speed and better characteristics the results might have been more accurate.

The time invested in this study is approximately of 340 hours, which have been divided into different tasks, from the initial planning until writing the final report. Furthermore, the budget used for this investigation is of 3.713€, which comprises the costs of the personal whom have work throughout the investigation and of all of the material used.

FUTURE WORK

Succeeding the conclusions of this investigation, several research paths for the future are proposed:

- A joint study of *valence* and *arousal*.
- A study of the images included in the video, along with a series of experiments in which acoustic characteristics studied in this project are combined with visual characteristics obtained in future researches.
- The creation of a segmentation tool, which takes into account the diverse content of the audios obtained from the LIRIS-ACCEDE database.
- An improvement of the reference file, which has been used in order to evaluate the segmentation. In order to do so, more videos must be classified and each video must be tagged by several people.

- Further study of the affective response based on the continuous analysis of the affective content of a video.